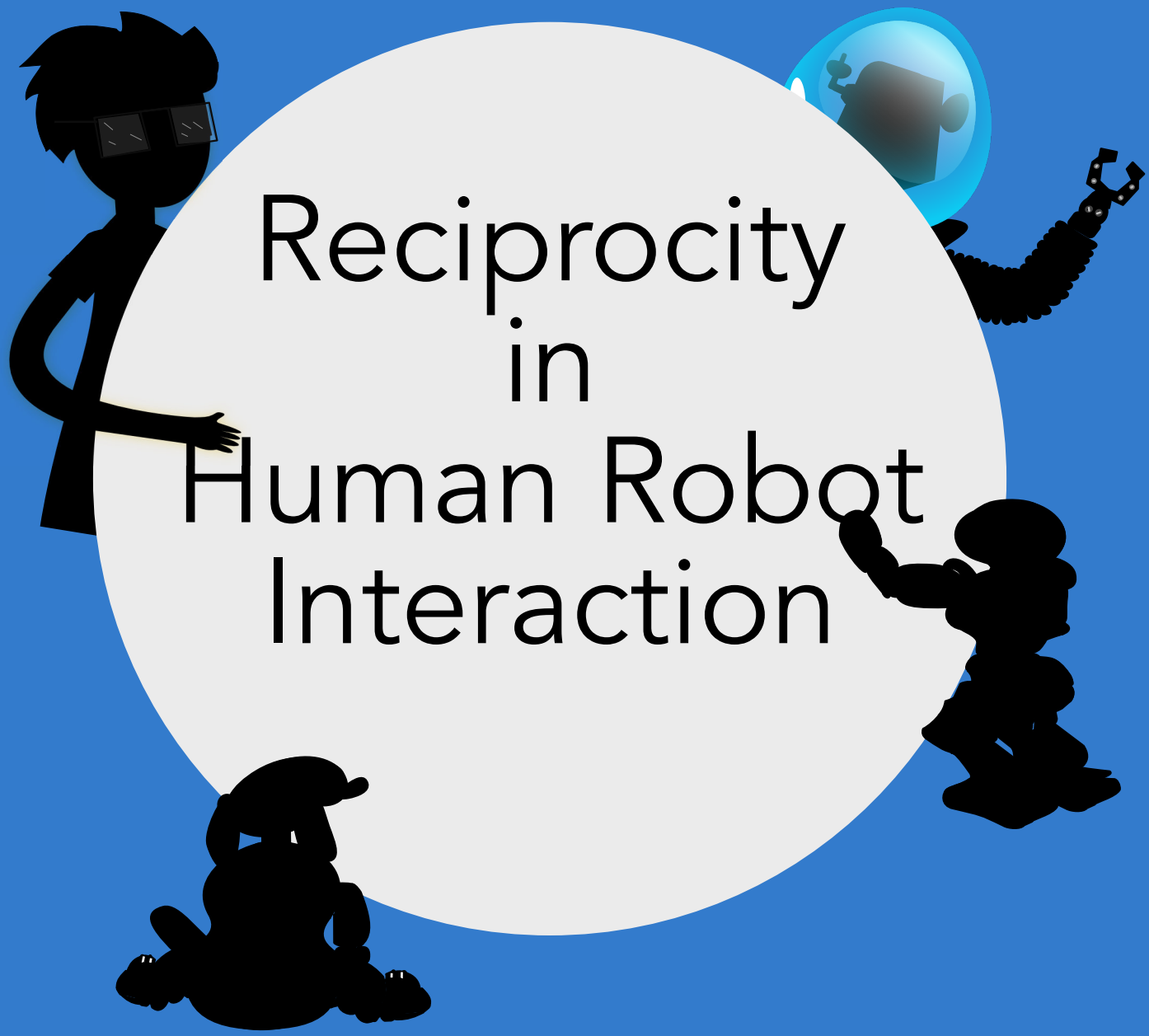


Reciprocity in Human Robot Interaction

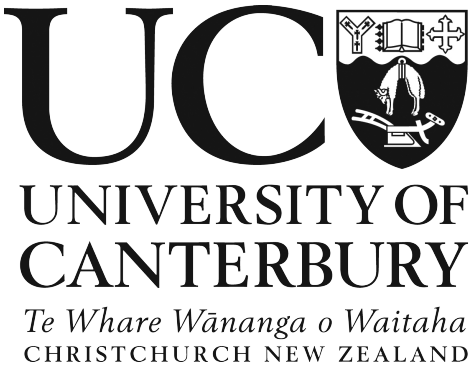
by
Eduardo
Benítez
Sandoval

Dissertation submitted
for the degree of
Doctor of Philosophy
in Human Interface Technology

UC
UNIVERSITY OF
CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Supervised by: Dr. Christoph Bartneck and Dr. Mark Billinghurst
Examiners: Dr. Takayuki Kanda and Dr. Selma Sabanovic

Reciprocity in Human Robot Interaction



Eduardo Benítez Sandoval

Human Interface Technology Laboratory New Zealand

This dissertation is submitted for the degree of

Doctor of Philosophy
in
Human Interface Technology

Supervised by:
Dr. Christoph Bartneck
Dr. Mark Billinghurst

November 2016

I would like to dedicate this thesis to my first teachers Inés, Delfina, and Juan, ...

Declaration

I hereby declare that the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This thesis is the result of my own work. Most of the content is the outcome of work already published as full papers in conferences and journals. I leaded led this research by myself in collaboration with other colleagues and my supervisor when was required. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations, and has fewer than 150 figures.

Eduardo Benítez Sandoval
November 2016

Acknowledgements

I would like to deeply thank my senior supervisor, Dr. Christoph Bartneck, for his support during my PhD journey. I am very grateful to Christoph for giving me the opportunity to come to New Zealand. This PhD has been the most challenging experience of my life, and I really appreciate his advice during its most complicated. Also thanks for his enormous patience, the extra work and the Eureka! moments during my research. Further thanks go to my co-supervisor, Professor Mark Billinghurst, for his suggestions regarding my first papers, his advice over the development of my career and his good example as head of the HITLab NZ. In addition, thanks to Dr. Bruce McDonald in the University of Auckland and Dr. Ricardo Sosa in AUT to allow me to work in their labs in the final stage of my PhD.

I sincerely thank my uncountable Kiwi, Mexican and international friends and their families in Christchurch and Auckland who become my family. Special thanks to Samuel, Koko and Hinoki-chan (and recently Keyaki-chan) and the rest of the whanau (Felicity, Clemie, Grant, and the others) Juergen and Latifah, Timo and Yuki, David and Grace, Alfredo Jimenez and Magnolia, Mohammad Obaid, Kirsten and Linda Erikson, Amit Barde, Huidong Bai, Hyungon Kim, Kuba, Seungwon Kim, Luis Nahmad, Pablo Lepe, Ricardo Sosa, Guillermo Ramirez, Omar Mubin, Victor, Alaeddin and their families. Also thanks to my many other friends in the HITLab NZ, I don't have enough space to mention all of you but you know very well how much I appreciate your friendship. Also thanks to the staff of the HITLab NZ, Ken, Lucia, Katy, Greg, Gun, Adrian, and more recently Rob; you do a wonderful job supporting the students.

Thanks to my family who raised me in the best possible conditions. I feel very lucky because I have a wonderful mother, Inés Sandoval, who supports me all the time and challenges me to be a better person. Her devotion to work, and her full dedication to help to her students and community have inspired me during all my life. Also thanks to my brother Rafael, my cousins, uncles and aunts, and the rest of my family to share their lives with me. Besides, I am very grateful to Fabiola, my partner. She is a very brave and patient woman who has shared this adventure with me. Fabiola, I love you so much.

I believe the individual success doesn't exist, it is just a product of our interactions with teachers, friends, family and community. Hence, thanks to my teachers along my life from

kindergarten to the PhD. Similarly, thanks to the public educative systems of México and New Zealand, specially to the IPN-UPIITA, UNAM-PDI, UAM-A, University of Canterbury and University of Osaka; my research is possible because these institutions do an excellent work with people like me.

This research was funded by NEC NZ Ltd. and I am very thankful to them; especially Glen Cameron. Thanks to Hamish House and David Humm in NZi3 and UC-ICT Business Development for their support along my research. I am very grateful to UC Doctoral and CONACYT scholarships to believe in me and support my research. I hope my research contributes a little bit towards making a better world through robotics technology.

Finally, thanks to to the robots of the HITLab NZ: Ikram, Socrates, Aristotle, Confucius, Kant, Schopenhauer, Papero, InMoov and Lego robot. We have had a magnificent reciprocal interaction along my PhD. I learnt a lot about humans with you guys.

Abstract

The Journey is the Reward
Chinese Proverb

Reciprocity is a basic characteristic of Human-Human Interaction (HHI). However, there have not been many previous studies about reciprocity in Human-Robot Interaction (HRI). The imminent coming of social robots interacting with users in their daily spaces has encouraged researchers in HRI to describe these new relations between humans and robots in terms of reciprocity, persuasion, likeability, and trust. Consequently, these studies could have an impact on the design of new social robots.

The development of this thesis considers three main research questions:

1. To what extent do humans reciprocate towards robots?
2. To what extent can robots use reciprocation for their own benefit?
3. What are the most beneficial and preferred reciprocal strategies between humans and robots?

I used Game Theory to develop three experimental studies. Decision games such as Prisoner's Dilemma, Ultimatum Game, Repeated Ultimatum Game and Rock, Paper, Scissors were used in the experiments. These games offer an engaging interaction between the participants and the robots, and they allow measuring of the variables related to reciprocity in HRI. The operationalisation of the studies was done under the definition of Reciprocity proposed by Fehr and Gaechter [53] which explains: *in response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self-interest model; conversely, in response to hostile actions they are frequently much more nasty and even brutal*. In addition to this, in the first and third study the "tit for tat" strategy was used with different modifications since it is a well-studied reciprocal strategy tested in previous experiments. Our main goal was to measure to what extent the Norm of Reciprocity,*"to those who help us, we should return help, not harm"* proposed by Gouldner [63] applies to Human-Robot Interaction in all the studies.

In the first study, we investigated whether reciprocal behaviours exist in Human-Robot Interaction and to what extent people reciprocate towards robots compared with humans. We designed an experiment that required participants to play the Prisoner's Dilemma game and Ultimatum Game with a NAO robot. We measured the number of reciprocations and collaborations between Humans and Robots and compared these with Human-Human Interactions.

In the second study, we investigated the negative side of the reciprocal phenomena in HRI to explore whether robots could use the natural human reciprocal response for their own benefit. In this study, we tried to answer questions such as: Can a robot bribe a human?

In the third study, we analysed the participants' preferences of the reciprocal robotic strategies. Since robots have identical physical embodiment, the design of appropriate robot-behaviours is very important as reciprocity plays a main role in the interaction between humans and robots. Our general research question in this study is: What type of robot behaviour is preferred by humans when the robot's decisions affect them?

On one hand people tend to conform to the Norm of Reciprocity in Human-Robot Interaction as they do with Human-Human Interaction but to a lesser extent; while on the other, humans find the unpredictable behaviour of the briber robots likeable and don't judge them in moral terms. They do, however, tend to reciprocate less towards robots who try to take advantage of the situation or show an unpredictable behaviour than with a robot that shows honest forward reciprocal behaviour. Furthermore, people prefer the most reciprocal and altruistic strategies of the robots compared with the selfish and most unpredictable reciprocal strategies. In other words, the construct of fairness in the form of reciprocity is present in HRI. In the future, once the robots have achieved an acceptable level of social skills our studies could be used as guidelines by robot behavioural designers.

Table of contents

List of figures

List of tables

Chapter 1

Introduction to Reciprocity

We are enriched by our reciprocate differences
Paul Valery

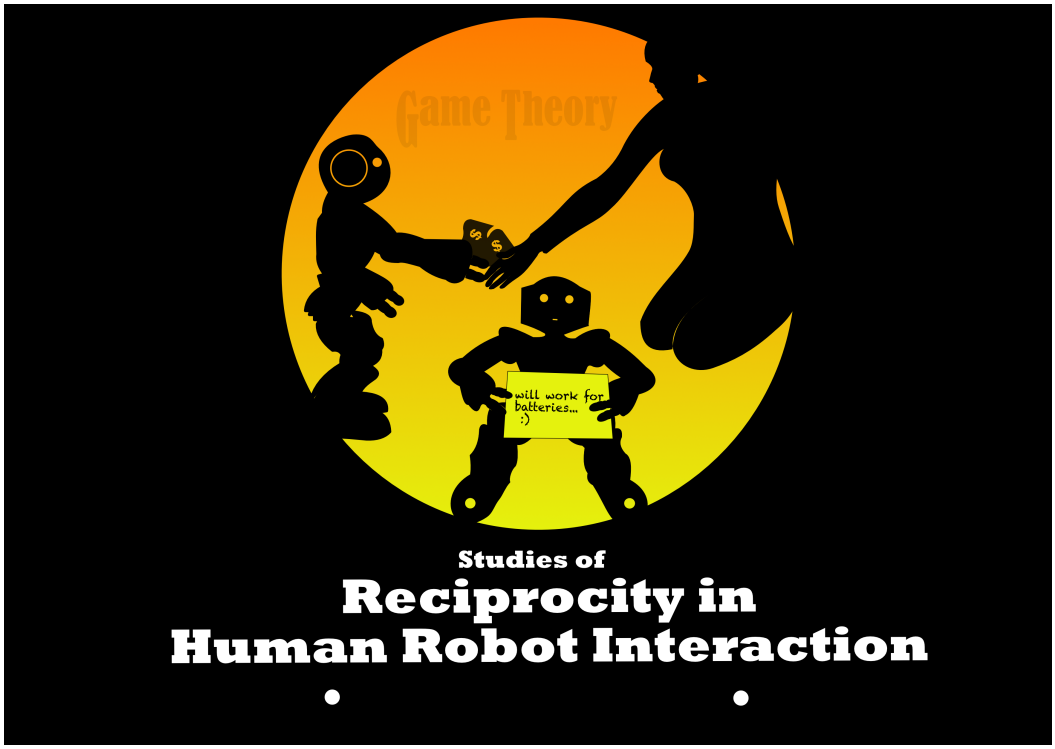


Fig. 1.1 Art developed for a presentation.

1.1 Structure of this Thesis

This thesis is intentionally designed as discrete sections, aimed at a very broad set of readers without the requirement for specialised knowledge in Human-Robot Interaction, Social Psychology, or Game Theory. However, my main goal was to write a consistent thesis with a linear development along our three main research questions. I hope that my thesis makes a positive contribution to the constantly evolving body of *HRI* research, and that *HRI* specialists are able to use our results for their own findings.

Most of the material in the thesis has already been published in Journals and Conference proceedings. Two main papers were published with novel material related to Reciprocity in *HRI* and are the product of the experimental work carried out for this PhD. The first paper *Reciprocity in Human-Robot Interaction.A Quantitative Approach Through The Prisoner's Dilemma And The Ultimatum Game* was published in the International Journal of Social Robotics [137] and sections of the paper are used in Chapter 1 and 2. My second paper, *Can a Robot Bribe a Human? The Measurement of the Negative Side of Reciprocity in Human Robot Interaction* [136] was published in the Proceedings of the International Conference of Human-Robot Interaction 2016 and is used in Chapter 33. Finally my third paper *Measurement of Reciprocal Strategies in Human Robot Interaction and their likeability using the Alternated Repeated Ultimatum Game.* was submitted to the proceedings of the International Human-Agent Interaction Conference in 2016 [138].

Several other research outcomes were developed during my PhD. All of them can be found in the appendix of this thesis or in www.sandoval.nz.

I hope you enjoy reading this thesis.

1.2 Historical Review of Reciprocity

Reciprocity has been studied for a long time in Philosophy, Ethics, Theology and Law. More recently it has been studied by Social Sciences, and Economics due to its importance in Human-Human Interaction (*HHI*). These days Reciprocity is an interesting topic in interdisciplinary fields like Behavioural Economics, Human Computer Interaction (*HCI*), and Interaction Design (*UX*).

Throughout human history, the importance of reciprocity has been a dominant feature of social interactions. Hence, many descriptions of reciprocity have evolved since the ancient times. For instance, the *Lex Talionis*: "An eye for an eye" [148], was proclaimed in the Hammurabi Code and it was an essential part of the Babylonian Law. Similarly, the Egyptians, Chinese, and Indians explicitly defined reciprocity in their moral codes (See Table 1.1). Due

to contact with other cultures, the ancient Romans enshrined these concepts of reciprocity within the principle *do ut des*: "I give that you might give".

Later, in the Bible we find one of the main references to the ethics of reciprocity still operative in Western culture, the *Golden Rule*: "One should treat others as one would like others to treat oneself". Indeed, the Bible makes emphasis of reciprocity and altruism in different passages such as Matthew 7:12 [100]: "So in everything, do to others what you would have them do to you, for this sums up the Law and the Prophets" and Luke 6:35: "...But be loving to those who are against you and do them good, and give them your money, not giving up hope, and your reward will be great and you will be the sons of the Most High: for he is kind to evil men, and to those who have hard hearts" [93]. The role of reciprocity is so important in the human moral life that 143 religious leaders claim the Golden Rule or Ethic of Reciprocity as a main common principle in the Declaration towards a Global Ethic [123].

In addition to the religious references, philosophers such as Epicurus offered a practical moral concept of reciprocity. The Epicurean Ethics of Reciprocity say: "Minimise the damage, the few and the many, to maximise the happiness of all" referring to the way that reciprocal justice should operate among human beings [24]. More recently Gouldner [63]claimed the *Norm of Reciprocity* as a universal human behaviour in 1960, saying: "to those who help us, we should return help, not harm" [20]. For Gouldner, this norm operates right across daily life. However, Gouldner's claim lacked controlled and quantitative experimental conditions so its application in experimental sessions has become an interesting area for research [109].

In recent years researchers have investigated reciprocity's material benefits, structure and motivations in order to support Gouldner's proposed Norm of Reciprocity. Many of these studies have taken a social, psychological, or anthropological approach. For instance, in 1974 Sahlins published his studies about reciprocity in terms of Economic Anthropology. He performed some qualitative observations unrelated to occidental economic practices, of non-western peoples including New Zealand Maoris. Sahlins [134] defines three kinds of Reciprocity (See Image 1.2.) as:

- *Generalised reciprocity*: Transaction that involves altruism, where the compensation does not have to occur in the short term, and cannot be paid. Generalised reciprocity is part of mutual aid among relatives, without expectation of material reward. The obligation to reciprocate is indefinite in time, quantity and quality.

- *Balanced reciprocity*: Direct exchanges based on certain equivalence with immediate restitution. Examples can include marriage arrangements, peace agreements or barter

of food products, as documented by Malinowski in Kula, Papua New Guinea. The remuneration must be within a defined time. Perfect Balanced reciprocity is given simultaneously and using the same types of goods.

- *Negative reciprocity*: A relationship in which a profit is intended at the expense of the other party with impunity. Included in this description is bargaining, cheating and stealing. Negative reciprocity can occur when participants have a social structural relationship, opposing interests and the aim to maximise profit.

Sahlins proposes that reciprocity, in its different modalities, is conditioned to the kinship between the participants. He considers that general reciprocity appears with the closest kinships and negative reciprocity appears in the more distant relationships. However, it should be considered that Sahlin's observations on the field were performed with specific human groups and these observations could differ from the relationships developed in contemporary societies with more extended social networks.

A different concept of reciprocity has been proposed by Malinowski [95] who focused his anthropological research on the exchange of goods among persons. Malinowski claims that people have non-altruistic motives for giving a gift, and expect an equal or greater value gift after some time. In other words reciprocity is an implicit process of gifting. These claims about the exchange of gifts generated a new research approach in Economics called Behavioural Economics, that investigates social phenomena in terms of material benefit. This topic is discussed in section 1.2.1.

Besides anthropological research, Social Psychology researchers also carry out experimental research on Reciprocity. Their experiments aim to validate the concept of reciprocity and offer data about variables like emotions and personality to describe what happens in reciprocal scenarios. For instance Cialdini [27]'s, approach to reciprocity through his study of Influence, Cialdini claims that it is possible to define a rule of reciprocation establishing that "we should try to repay, in kind, what another person provided us" [27]. Further, he defined some universal principles whereby people have influence over others, in which reciprocation plays a key role. These include authority, trust in experts, commitment/consistency (people acting consistently with their beliefs), scarcity (people more intensely desiring less available resources), liking (people tending to say yes to other people who like them), and social proof (people looking to the behaviour of others to guide their own) [59]. Other authors have analysed additional factors through experiments about the changes in reciprocity, for instance Whatley et al. [158] investigated reciprocal scenarios such as when a favour is asked in a public or private circumstance; an experiment that is possible to reproduce with robots.

Table 1.1 Ancient references of Reciprocity in different cultures.

Culture	Definition
Christian/Catholic Golden rule	One should treat others as one would like others to treat oneself
Judaism Leviticus (19:18)	You shall not take vengeance or bear a grudge against your kinsfolk. Love your neighbour as yourself .
Judaism Talmud (Shabbat31a)	That which is hateful to you, do not do to your fellow.
Romans Do ut des	I give that you might give
Romans Stoicism (Seneca)	Treat your inferior as you would wish your superior to treat you
Egyptians Ma'at	Now this is the command: Do to the doer to make him do
Indian Sanskrit Tradition	...by self-control and by making dharma your main focus, treat others as you treat yourself.
Indian Tamil Tradition	Why does a man inflict upon other creatures those sufferings, which he has found by experience are sufferings to himself ?
Chinese Confucius	Never impose on others what you would not choose for yourself
Greeks Thales	Avoid doing what you would blame others for doing
Greek Isocrates	Do not do to others that which angers you when they do it to you
Islam Quran Muhammad	Wish for your brother, what you wish for yourself
Hinduism	One should never do that to another which one regards as injurious to one's own self. This, in brief, is the rule of dharma. Other behaviour is due to selfish desires.
Buddhism Buddha	Just as I am so are they, just as they are so am I
Jainism Sutrakritanga	A man should wander about treating all creatures as he himself would be treated.
Taoism	Regard your neighbour's gain as your own gain, and your neighbour's loss as your own loss

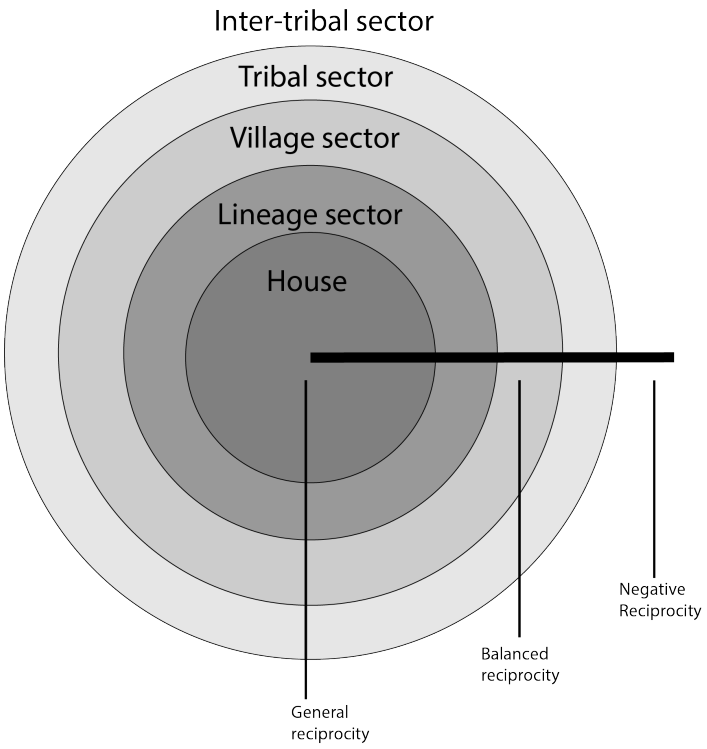


Fig. 1.2 Reciprocity in terms of kinship distant relationships, (Sahlins, 1974.)

Some other explanations about the motivations in the reciprocal process come from Evolutionary Biology. In the 1970s, Trivers [150] presented a model referred to as "reciprocally altruistic behaviour". This model describes the behaviour of an organism with apparently detrimental attitudes towards an organism that is not closely related, but expecting that the other organism will act in the same reciprocal way later. This model describes reciprocity considering two main responses involved in the process: reciprocation and no reciprocation, and proposed the consequences of both acts. In the first case, reciprocation generated a reinforcement of friendship. In the case of no reciprocation indignation and anger towards the organism that was not reciprocal resulted. However, this model is a simplification of the reciprocity process that does not take into account the number of favours, attempts to be reciprocal, or the initial intentionality. See Image 1.3.

As we can see, there are at least three well defined approaches of the studies of reciprocity: the social-exchange, the social norms, and the evolutionary approach. These approaches mainly try to explain the motivations and collective benefits of the actors in reciprocal scenarios. Myers summarises these different research approaches in 1.2.

Besides the investigation related to the description and motivations of reciprocity, other researchers in the Economic field try to measure the reciprocity in order to know the internal mechanisms behind the not always rational decisions taken in HHI. Research in Economics

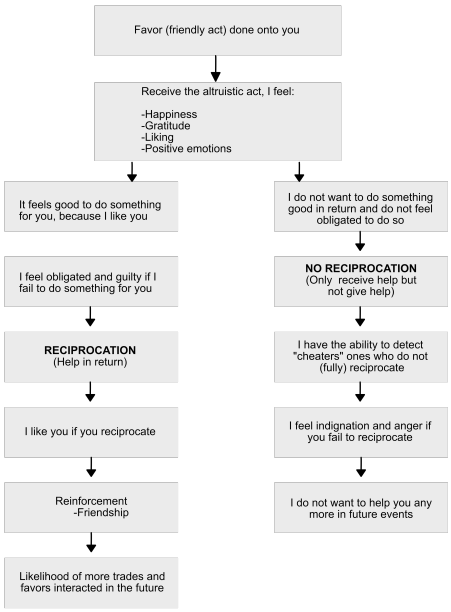


Fig. 1.3 Evolutionary Biological approach of the process of reciprocation (Trivers, 1973)

Table 1.2 Summary of the Theories explaining reciprocity and their motivations.

Theory	Level of explanation	Externally Rewarded Helping	Intrinsic Helping
Social-exchange	Psychological	External rewards for helping	Distress-rewards for helping
Social norms	Sociological	Reciprocity norm	Social responsibility norm
Evolutionary	Biological	Reciprocity	Kin selection

also investigates the implications of reciprocity in fairness, altruism, friendship, love and other social phenomena.

In section 1.2.1 I discuss the definitions of reciprocity mainly based in quantitative approaches of reciprocity that are the baseline of the methods used in this research.

1.2.1 Definitions of Reciprocity

The progress of the research in reciprocity has recently taken a scientific approach. Aside from the pure moral and religious definitions of reciprocity new and more precise definitions have been proposed coming from observations and experimental work. For instance, the Encyclopaedia of Anthropology claims that Reciprocity “is the state of mutually addressing the same attitudes or feelings as another. It indicates an equal exchange...in a world where there is no external authority to enforce agreements” [12, 121]. Similarly, Kolm [81] defines reciprocity as “a set of motivational interrelated gifts or favours” [81, 82].

Along the observational definitions in Anthropology, there are definitions of reciprocity in Economics validated by experimental work in controlled conditions. Indeed, there is a large corpus of research related to the measure of the effects of reciprocity in altruism, friendships and different kinds of relationships edited by Arrow and Intriligator [4] who published the *Handbook of Economics of Giving, Altruism and Reciprocity* focused totally in Behavioural Economics. Furthermore, Falk and Fischbacher [50] propose a Theory of Reciprocity based in experimental work to test these concepts. The basic concept that they developed is *"that a reciprocal action is modelled as the behavioural response to an action that is perceived as either kind or unkind"*.

Besides the theory of reciprocity, Fehr and Gaechter [53] claims that *"Reciprocity means that in response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self-interest model; conversely, in response to hostile actions they are frequently much more nasty and even brutal"*. This definition will be consider for the operationalisation of studies one and three, because:

- It is validated by experimental work and it matches with the scope of this research.
- It covers a full spectrum of reciprocal phenomena from the positive reciprocal acts to the negative reciprocal acts.
- It has been used previously in *HHI*; hence, it can facilitate compare *HHI* studies vs *HRI* studies.

I consider that the definition proposed by Fehr and Gaechter [53] is sufficiently specific for my experiments while at the same time being optimal in terms of implementation to be used as baseline of this research. Furthermore, it is in line with the theory of reciprocity proposed by Falk and Fischbacher [50] based on experimental work. The theory explains a reciprocal action modelled as the behavioural response to an action that is perceived as either kind or unkind. Furthermore, it is in line with the work of Nass and Moon [110] who conclude that in the context of interaction with machines as computers; *"One should provide help, favours, and benefits to those who have previously helped them"*. The definition of reciprocity proposed by Fehr and Gaechter [53] is in line with some experiments in Social Psychology that show possible differentiation between different motivational suppositions in the reciprocity process [52].

1.3 The study of Reciprocity in HCI and HRI

This an edited version of sections of the paper Reciprocity in Human-Robot Interaction: A Quantitative Approach through the Prisoner's Dilemma and the Ultimatum Game, [137].

Goodrich and Schultz [60] defines *HRI* as: *"... a field of study dedicated to understanding, designing and evaluating robotics system for use by or with humans".* By this definition, it is possible to infer the communication between humans and robots and assume that this communication could be social and very sophisticated. Indeed, humanoid/social robots have the proper morphology and sensing modalities that make them a technology particularly suited to these purposes. Furthermore, Breazeal [20] claims that the interaction of the robots with humans would generate new kinds of applications in homes, entertainment, and health care. These kinds of application for social robots could be naturally social and would motivate the future design of new robots capable of interacting and cooperating with people as colleagues in better way than with a computer, smart phone or similar technology.

Additionally, it is assumed that robots and other machines should be cooperative with humans but these studies have not considered reciprocity as a main factor in this phenomena. According to Cialdini [27] and Perugini et al. [124] cooperation between humans lies in reciprocation and persuasion. Furthermore, the studies of Nass and Reeves [112], Nass and Moon [110], and Fogg [58] suggest that similar cooperative phenomena could happen between humans and computers. In other words, if a person and a machine socially interact, this interaction implies reciprocity and persuasion in both directions. However, they don't explain to what extent all these social machines will reciprocate towards humans if they are compared to *HHI* similar scenarios. Fogg developed the concept of persuasive machines [56–58], considering that humans have an instinctive behaviour towards devices that triggers feelings and emotions in response to their persuasiveness. These feelings and emotions are apparently reciprocal to the machines when they provide a good service or help. In other words, *If you are nice to me, in the future I will be nice to you.* Indeed, in the early 1990s Fogg and Nass ran a small experiment that tried to demonstrates that users tend to return a favour with computers that had helped them previously [55].

Besides, negotiation is an activity which inherently involves reciprocity in order to obtain satisfactory results for negotiators. Several studies have been done with automated agents negotiating in different decision scenarios. Lin and Kraus [89] offer an extensive review of these agents in [89]. The performance of the agents varies statistically significantly depending on the scenario and the internal design of the algorithms. Moreover, Kiesler et al. showed that humans show cooperative behaviour towards computers [79] playing Prisoner's Dilemma when they have a chance to interact intensively with the agent. In this Prisoner's Dilemma

Table 1.3 Primary types of social cues

Primary Types of Social Cues, factors related to the interaction with persuasive technology	
Cue	Examples
Physical	Face, eyes, body, movement
Psychological	Preferences, humour, personality, feelings, empathy, "I am sorry"
Language	Interactive language use, spoken language, language recognition
Social dynamics	Turn taking, cooperation, praise for good work, answering questions, reciprocity
Social roles	Doctor, teammate, opponent, teacher, per, guide

the cooperation is conditioned to the previous actions of the other participants; if a player was cooperative or defective that could condition the response of the opponent in the next round (reciprocal behaviour), so *I will be nice with you now because in the future I expect that you will be nice to me too*. I will explain Prisoner's Dilemma in detail in Chapter 2. Besides, De Melo et al. also used Moral Emotions (gratitude, anger, reproach, sadness) to elicit cooperation with a virtual agent in 25 rounds of Prisoner's Dilemma using a variety of Tit for Tat strategy [102]. However none of these studies consider reciprocity as a main variable to be measured.

The studies previously mentioned explain that *HCI* happens due to the human tendency to anthropomorphize objects. The Media Equation is a widely accepted theory in *HCI* proposed by Nass and Reeves [112]. The theory suggests that people tend to treat certain objects in the same way that they treat other people. Similarly, Norman also have conducted research supporting the idea that people interact with objects and tend to attribute human features and project their feelings over them [111, 115–117]. In other words, people are fundamentally social and natural with computers, television and other media such as smartphones and maybe robots.

The expectation of many the researchers in Human-Robot Interaction (*HRI*) is that robots will continuously communicate with people in everyday life. Ideally, robots will take their own role in society and they so must persuade people to trust them and create a perception of reliability. However, we require more information to claim that is possible. Human behaviour is complex and people develop intricate relationships with other humans, pets, machines and objects in their lives. Since *HRI* often mimics *HHI*, it is expected that reciprocity will also play an important role in *HRI*.

Science fiction and popular TV shows present future hypothetical relationships between humans and robots and in particular the problems around reciprocity. Although, these situations are fictional, they offer a reference for the development of reciprocal strategies in the design of robots [8]. For instance, the TV Show *The Simpsons* chapter *Them, robots* offers an ongoing example of human frustration (due to the lack of reciprocity) in a workplace

where both humans and robots are collaborating and competing simultaneously to achieve success in the development of a group task [126]. Similarly, *Futurama* continuously shows a robot with a selfish personality and abusive attitude toward humans [99]. By contrast, the movie Robot and Frank is a more lifelike situation; a relationship between an old man and a robot facing ethical dilemmas when the robot is trying to persuade Frank to engage in a more healthy lifestyle and Frank's reciprocal attitude is very negative [140].

Similarly to Science-Fiction we expect that many social situations in the future will involve *HRI* related to reciprocity. This is particularly the case for service robots that act within the legal framework required for their operation [47]. There are different dilemmas to solve in this area. For example, racism and classism in human societies throughout history have been based on a lack of reciprocity, causing phenomena like slavery and discrimination. These relationships were not reciprocal and not fair among masters and slaves. Could the same happen with robots? Would robots be ordered to wait outside of the public places such as pubs or restaurants for their masters? Would robots be treated as our equals? Would we be reciprocal with them? Would we pay for their services and take care of them? These questions are out of the focus of this research but they are matter for future studies.

1.3.1 Previous Studies of Reciprocity in HRI

Several studies related indirectly to reciprocity in *HRI* have been done in the domain of Companion robots. Companion robots are a subset of social robots and service robots which will become popular in the near future. Dautenhahn et al. [36] described them as robots designed for personal use; capable of performing multiple tasks and interacting with the users in an intuitive way. Studies in social robotics propose the use of these robots in different scenarios. For example as educators, caregivers in nursery houses, nannies, housekeepers, and assistants. In fact, important research consortia like The Cognitive Robot Companion [1] and Robot Companions for Citizens [2] are investing resources in the development of companion robots. Moreover, it is expected that users and robots develop short-term and long-term relationships if companion robots assume certain social roles in the life of the users. In *HRI*, it is commonly used Human-Human Interaction (HHI) as a reference to compare our robotic implementations.

Kahn et al., [76] discovered that children responded reciprocally and were more engaged with an AIBO robot which offered some motioning, behavioural and verbal stimulus than they were with a toy dog. More recently reciprocity has been very present in the debate of social robotics. The workshop: *Taking care of Each Other: Synchronisation and Reciprocity for*

[1] www.cogniron.org
[2] www.robotcompanions.eu

Social Companion Robots in the International Conference of Social Robotics 2013 discussed the importance of reciprocity in the design of companion robots. Several studies presented in the workshop reviewed concepts related to reciprocity as compassion, behaviour imitation or social cognition mechanisms integrated to *HRI* [156] which could be the cornerstone in the development of future meaningful Human-Robot Interactions.

For instance Weiss presented the project Hobbit [155], a robot based in the *Mutual Care* paradigm proposed by Lammer et al. [87]. They, like me, propose that Human-Robot Interactions can improve if both parties take care of each other in a similar way to human human interactions. Furthermore, Lorenz claims that mutual compassion (understanding Compassion as the German word *Mitgefühl* should be considered as an important component in *HRI* due to this being a human ability based in synchronisation and reciprocity. The benefits of mutual understanding based in a reciprocal relationship between humans and robots can improve the performance of social companion robots because of the resulting more intuitive behaviour of the robot [92]. However, Broz and Lehmann claim that reciprocity is limited to certain *HRI* scenarios where robots assist humans in some activities and humans assist the robots in others. Although cooperation and reciprocity are closely related, they do not necessarily appear together. For instance in jobs as caregivers, which could be likely future roles for companion robots, the patients do not necessarily behave in a reciprocal manner with the caregiver [22].This could be because the robot doesn't encourage reciprocal behaviour, but it is likely that this lack of reciprocity in *HRI* can produce a depreciation of the services provided by the robot and consequently the construction of a relationship will be degraded. We think that reciprocity is especially important if the users need the robot. If something happens to the robot but the user does not care, the user will suffer a negative impact later because the robot could not do its work. In our opinion other roles for companion robots could require a more reciprocal behaviour when a social interaction is developed.

Additionally, decision games have been used to study different aspects of *HRI* [142] to study reciprocity indirectly. To illustrate, Nishio et al. [113] have studied how the appearance of robots affects participants in an Ultimatum Game (UG). This game involves reciprocity because two players interact to decide how they will divide money or points in fair or unfair proposals. Nishio et al. conclude that people show changes in their attitude depending on the agent's appearance. The agent (robot, human or computer) in the role of proposer influences the number of the rejections of the proposals. In particular an android appearance is associated with a higher number of rejections. Possibly not enough human likeness in the android's appearance is a main factor. In addition, Torta et al. used Ultimatum Game online to measure the perceived degree of anthropomorphism among a human agent, a humanoid

robot and a computer. In that study, participants took more time to respond to the offer of a computer compared to that of the robot [149].

An analysis of reciprocity in *HRI* could be useful in order to design more engaging and effective Human-Robot Interactions in different scenarios. Some studies report that users do not feel engaged enough with the robots and that they have high initial expectations of them which decrease over time [21, 46]. On the other hand, companion robots have not so far had the expected impact in people's lives, particularly when they take care of particular users such as elderly people [21] or children. Dautenhahn et al. found that 40% of the users liked the idea of a companion robot in the home. In addition 96.4% of the users wanted a robot capable of doing the housework. However a robot playing a role in the human domain as friend or taking care of children was acceptable to only 18% of the participants [36]. We propose that in the future, robots could assume more social roles in the human domain if the Human-Robot Interaction would be more reciprocal.

1.3.2 Future Design of Reciprocal Robots

Reciprocity is a critical factor in *HRI* because robots are designed assuming that they should be collaborative with humans in order to do different task together. However, humans differ widely due to factors as personality, culture, age or gender. Consequently, some humans would accept robots more easily than others, even though the robot shows highly collaborative and reciprocal behaviour with all of them. This is critical for some future applications. For instance, what kind of old man could be assigned a robot to look after him? What kind of child would be a good candidate to be cared for by a robot? What attributes should an adult personality have to work effectively with a robot? Moreover, how should robots be designed to match with these requirements? It is also important to notice that collaborative behaviour is not exactly the same as reciprocal behaviour. The latter refers to a mirroring type of behaviour (e.g. an eye for an eye) whereas collaborative behaviour usually refers to only the positive form of reciprocal behaviour. My research questions aim to sort some of these issues in the future.

Reciprocity could involve either collaboration or retaliation. Hence, it is important to investigate Reciprocity in *HRI* in order to establish the parameters for the design of robots capable of being reciprocal and persuasive with humans in a positive way. Moreover, it could be extremely useful; according to our research questions, to know to what extent people reciprocate (positively and negatively) towards robots, if robots can use reciprocity in their own benefit and what are the most efficient robot reciprocal strategies in order to obtain the maxim mutual benefit in the *HRI*. It is evident that in the near future, robots will have technical limitations and limited social skills. Considering that reciprocity is one of the most

important phenomena in human social activities, it can be assumed that this should be studied in order to know how reciprocity influences the long-term relationships of humans and robots. In addition, it should be considered that humans have a high capability to adapt to agents when they are interacting with them.

Despite the importance of reciprocity in *HCI*and *HRI*, the area has still not been explored enough. The research related with reciprocity is mainly focused in persuasion, negotiation and cooperation. Apparently the community of social robotics accepts reciprocity as a fact. However I consider that reciprocity should be measured and compared in order to have a reference to be used as a guideline in the design of new interactions.

A main problem is that there are not many previous works about reciprocity in *HRI* due to the technical limitations that robots had interacting actively with humans until this time. The imminent coming of robots interacting with users in the daily life suggests that it is necessary describe this new relations based in factors like reciprocity, persuasion and trust. Moreover, reciprocity is connected with other phenomena such as persuasion [27] cooperation [7] , altruism [80], friendship [29], love [85] and compassion [155]. Hence, we need to know the impact of reciprocity in the design of new social robots. It is very likely that reciprocity will play a major role in *HRI* in the future.

In terms of applications using the reciprocity concept in *HRI*, there are examples of how the design of reciprocal behaviours could be applied with Children with Autism Spectrum Disorder [108] or elder care [87]. However a better understanding of reciprocity could help to improve the current use of companion robots in real applications like in the work presented by Broadbent et al., [21]. They report that the robots in their experiments did not have enough impact on the quality of life of the patients, mitigate their depression or create adherence. Hence, this study is an opportunity to consider that a more reciprocal behaviour of the robots could help to improve their performance with the patients.

Many of the interactions between humans and robots involving cooperation, persuasion, altruism, exchange of favours or mutual trust could depend on reciprocity. In addition, it should be considered that humans have a high capability to adapt to agents when they are interacting with them depending of their own personality traits. For instance, people could be reciprocal with a robot by paying it back for its services (taking care of the robot, giving technical maintenance, etc) if the robot encourages reciprocity via certain social strategy. Authors like Kahn et al. consider reciprocity as a benchmark in the design of Human-Robot Interactions [73] simply because reciprocity is present in other human social situations. Moreover, they raise an interesting question in *HRI*, which is: Can people engage substantively in reciprocal relationships with humanoids? They argue that interactions involving reciprocity with anthropomorphic robots can be similar to human interactions

[73]. In other words, humans tend to develop intricate relationships with pets, machines and artifacts, consequently, it is expected that reciprocity plays an important role in *HRI*. However, the question is; do people reciprocate towards robots in a similar way to how they reciprocate with humans?

1.4 Aims, Scope and Research Questions

In this *HRI* research I use a multidisciplinary approach coming mainly from Behavioural Economics, Social Psychology and Interaction Design. My approach is experimental and quantitative to discover in which ways and to what extent people reciprocates towards robots. Hence, I decided to use the insights of Game Theory in order to explore Reciprocity in *HRI*.

1.4.1 Game Theory

Game theory is a branch of Behavioural economics that defines the methods employed to define a decision's patterns in economical scenarios. Williams [159] define Game Theory as *an interdisciplinary theorist method that examines how people make decisions when their actions and fates depend on the actions of other people*. In this research I used different variants of decision games such as: Repeated Prisoners Dilemma (RPDG) and Ultimatum Game that can offer us a quantitative reference of reciprocity in *HRI*. These decision games will be explained in Section Chapter 2.

The Decision Games are a common research tool used in *HHI* to investigate other related phenomena as cooperation or negotiation as they allow a simplification of different social situations. Additionally these games can be changed to model other scenarios. For instance, Prisoner's Dilemma could be adjusted without modifying the essence of the game for different situations where participants should take decisions such as in wars, law enforcement, or duopoly fights [147]. Some examples of the use of Game Theory used in *HHI* can be found in [82, 86, 127, 159].

Game theory also offers different kinds of experiments: Cooperative vs. Non-Cooperative, Competitive vs. Non-competitive, Normal vs. Extensive games. The area of Game Theory that is useful for the development of a Model of Reciprocity is Cooperative Games. This kind of game involves the concept of reciprocity in its development. Cooperative games could be described as the kind of games in which *"players can communicate with each other and form binding coalitions and pacts, or agreements among members to coordinate any strategic action"*. [159]. A comprehensive review of Game Theory assumptions, application and limitations can be found in [128, 147, 159].

However, detractors of Game Theory argue that the outcomes of decision games are difficult to transfer into real situations where such variables as reciprocity are not under controlled conditions for several reason . For instance, If the number of players is increased the actual decision tree generated becomes more difficult to predict and analyse. Besides, Game Theory just provides the general logic for the game and not necessarily the winning strategy [128].Furthermore, the combination of factors and variables conditions the benefits in a reciprocal interchange of goods [159].

In contrast to this arguments, I propose that decision games are a suitable tool to perform exploratory studies, and most of the studies performed in *HRI* due to the current limitations of real interaction in controlled conditions. In addition, Game Theory is focused in the material outcomes of the economic scenarios, opposed to the motivations of the participants which is also interesting in term of Social Psychology. Hence, Game Theory is suitable to measure to what extent people and robots reciprocate. Then, I propose the use of decision games as validated methods to obtain measurable result able to be analysed statistically and explained by Social Psychology in terms of a multidisciplinary approach. Likewise, Social Psychology and Interaction Design have helped to model the experiments performed along this research.

I used the insights of Social Sociology and Interaction Design due to the extensive corpus of research in Social Psychology, Communication, and Interaction that describe how humans interact with other humans. *HHI* models have been used as a baseline to the development of similar models in *HCI*and *HRI*. Humans can use oral communication, body language, gazes or simple sounds. In addition, Human-Human Interaction involves factors such as: family ties, bonds of friendship and hierarchies which are strongly correlated with favour exchanges, social punishment and even ostracism connected with reciprocity. Humans also have interactions with other living beings like their pets and they tend to anthropomorphize and interact with in speech or other types of communication.

1.4.2 Research Questions

As has been reviewed, reciprocity is a basic characteristic of *HHI* . However, there have not been many studies about reciprocity in *HRI*. *HRI* researchers are motivated to describe the new relations between humans and robots in terms of reciprocity [53, 63], persuasion, likeability defined by Cillessen and Rose [28] in 4, and trust due to the development of better social technology Consequently, these our studies could have an impact in the design of new social robots. This thesis covers three main research questions:

1. To what extent do humans reciprocate towards robots?

2. To what extent can robots use reciprocation for their own benefit?

3. What are the most beneficial and likeable reciprocal strategies between humans and robots?

Chapters two, three and four address these questions in turn. The answers are subdivided into other research questions in order to simplify and design experiments that can contribute to the *HRI* state of art. We expect that an answer to these questions would be a significant contribution to *HRI* research in terms of robot behavioural design in the near future.

Chapter 2

Measurement of Reciprocity in Human Robot Interaction

Modern science is predicated on 'truths' verified through accurate observation and measurements of physical world phenomena.
Bruce Lipton

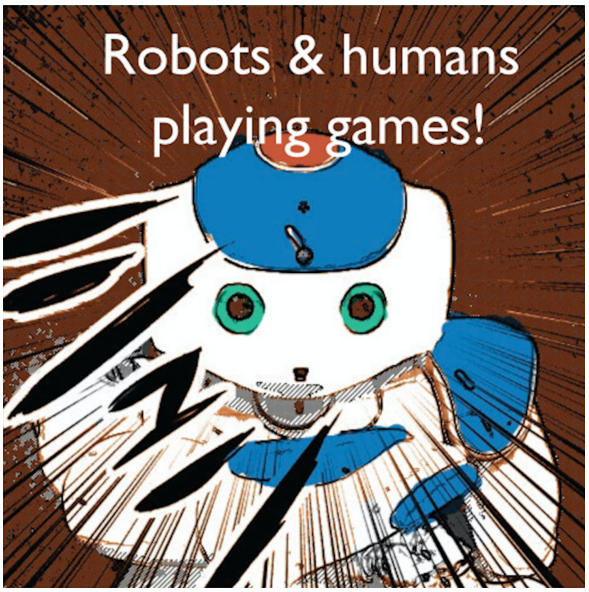


Fig. 2.1 Art developed for recruitment.

The material of this chapter was already published in: Sandoval, E. B., Brandstetter, J., Obaid, M., and Bartneck, C. (2016b). Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. International Journal of Social Robotics, 8(2):303–317. [137].

2.1 Summary

In this chapter we describe how participants in our study played the Repeated Prisoner's Dilemma game (RPDG) and the mini Ultimatum Game (mUG) with robot and human agents, with the agents using either Tit for Tat (TfT) or Random strategies. As part of the study we also measured the perceived personality traits in the agents using the TIPI test after every round of RPDG and mUG.

2.2 Introduction

2.2.1 Game Theory as a research tool in HRI

To explore reciprocity we decided to use the insights of Game Theory. We used Repeated Prisoner's Dilemma (RPDG) and Ultimatum Game as the decision games that could offer us a quantitative reference of reciprocity in HRI. Both games are a common research tool used to investigate other related phenomena as cooperation or negotiation allowing simplification of different social situations. Additionally these games can be changed to model other scenarios. For instance, Prisoner's Dilemma could be adjusted without modifying the essence of the game for different situations where participants should take decisions such as in wars, law enforcement, or duopoly fights [147].

2.2.2 Prisoner's Dilemma

The Prisoner's Dilemma game is frequently used as a quantitative approach to study different phenomena. Since Rapoport and Chammah proposed the Prisoner's Dilemma in 1965 [129] there have been different versions of the experiment which differ in the terms of the defection and collaboration required of the players. In the original game two thieves are captured by the police and interrogated separately. They can cooperate with each other keeping quiet or they can defect confessing the crime, but the punishment of both thieves depends of the combination of cooperations or defections of each. The rules are: "There are two players. Each has two choices, namely cooperate or defect. Each must make the choice without knowing what the other will do. No matter what the other does, defection yields a higher payoff than cooperation. The dilemma is that if both defect, both do worse than if both had cooperated" [7]. One of the matrix versions of the game is shown in Table 2.1 [147].

In Table 2.1 the numbers represent time in prison for the participants in the game. The minus sign is a convention to indicate that this time is subtracted from the time of the criminal in the metaphor. To illustrate, if Criminal 1 and Criminal 2 both cooperate (keep quiet), both

Table 2.1 Basic Prisoner's Dilemma Matrix

		Criminal 2	
		Cooperate	Defect
Criminal 1	Cooperate	(-3,-3)	(-12,0)
	Defect	(0,-12)	(-8,-8)

will spend just three months in jail. However, if Criminal 1 cooperates and Criminal 2 defects, Criminal 1 will spend 12 months in jail and Criminal 2 will be free. If both of them defect they will spend eight months in jail. The game represents situations where simultaneous decisions affect two parties.

Defect offers the highest profit for the players when the game is played once. Therefore strict dominance here is Defect. Spaniel defines a strict dominance when "We say that a strategy X strictly dominates strategy Y for a player if strategy X provides a greater payoff for that player than strategy Y regardless of what the other players do." [147]. In other words, when we have a strict dominant strategy in a decision game it should be clear for the participants what to decide in order to get the highest profit. For a single round of Prisoner's Dilemma, Defect is the strict dominance strategy because it allows a player to avoid punishment. However when many rounds are played, Cooperate or Defect are possible strategies to reduce the punishment of both players.

Diverse versions of Prisoner's Dilemma have been developed. For instance, Prisoner's Dilemma can also be played in consecutive rounds, which is called Repeated Prisoner's Dilemma Game (RPDG) modality. In this version, previous movements of the opponent become a factor for the next movement of the player, who is probably considering and recording the behaviour of his opponent [159]. Furthermore, about 20 strategies have been tested in order to get a good score in the RPDG [5, 38]. According to Axelrod the strategy designed by Rapoport, "Tit for Tat" (TfT) is the simplest and most effective strategy to follow in the RPDG [7]. Tit for Tat consists of cooperating in the first instance and then in the next movement copying the decision of the other participant did in the previous round. In two contests organized by Axelrod in the 1980s different strategies were tested. In both contests Tit for Tat was the winner [6]. As can be seen, this strategy has been well- studied in *HHI* and is regarded as a pattern of reciprocal strategies This strategy has been used in this study and also in the study described in Chapter 4 with certain modifications.

2.2.3 Ultimatum Game

In this game, one of the participants (Proposer) decides how to distribute a certain amount of money. The second player (Acceptor) can decide to accept the distribution and both of them

can keep the money. However, if the acceptor rejects the offer both of them lose the money. Like Prisoner's Dilemma, Ultimatum Game has different variants. One is the mini Ultimatum Game (mUG) in which participants decide upon a limited set of defined distributions of money, for example, 50%-50%, 20%-80%, 80%-20%, or other options [49]. For this study, we use the mUG version of the Ultimatum Game, and fixed the roles for the agent and the participant. Participant is always the proposer and Agent is the acceptor.

2.2.4 Studies of personality and reciprocity

Several researchers claim that human personality matters in games related to reciprocity such as Prisoner's Dilemma. Park et al. claim that the behaviour in situations involving reciprocity is affected by personality and the interactions of the parties following the norm of reciprocity. In addition, they suggest that extroversion, agreeableness and neuroticism personality traits are related to cooperative strategies in conflict resolution. [122]. Boone et al. conducted an experiment which deals with four personality traits: locus of control, self-monitoring, type-A behaviour and sensation seeking [16]. In addition, Chaudhuri et al., performed the Repeated Play Prisoner's Dilemma (RPPD) researching trusting and reciprocal behaviour [26]. They classified people with different propensities to cooperate showing differing degrees of trust and reciprocity. They found that people who chose to cooperate demonstrated higher levels of trust. In contrast, in reciprocal behaviour, differences between cooperative subjects and defectors were not significant.

2.3 Research Questions

Our general research questions for this study are: Do people reciprocate differently towards other humans in comparison to robots? What consequences does the interaction strategy of the robot have on the humans' reciprocal behaviour? In order to answer these questions we developed the following sub questions:

1. Do participants behave differently towards robots compared to other humans in terms of reciprocation, collaboration and the offer they make in the ultimatum game (Offer)?

2. Do participants behave differently towards agents that use the TfT strategy in comparison to how they behave with agents that use the Random strategy in terms of reciprocation, collaboration and the offer they make in the Ultimatum Game (Offer)?

3. Do participants win more money (Human Profit) when the agent uses the TfT strategy compared to when the agent uses the Random strategy?

4. Do participants and robots together win more money (Joint Profit) when the agent uses the TfT strategy compared to when the agent uses the Random strategy?

5. Is there any correlation between Collaboration, Reciprocation, Human Profit and Joint Profit?

6. Is the personality of the agent perceived differently when the agent uses the TfT strategy compared to when the agent is using the Random strategy and how is this relationship mediated by the participants' own personality?

2.4 Method

The aim of this study is to model reciprocity with a quantitative approach in order to understand the reciprocal actions of the participants towards the robots. We used the Repeated Prisoner's Dilemma Game (RPDG). The participants played ten rounds similar to the experiment of Selten and Stocker who did a series of "super games" playing 25 times in periods of ten rounds [141]. Then the participants played as proposer and the agent as acceptor in the mini Ultimatum Game (mUG).

In our study the participants did not know how many times they would play against the agent. That means that their decisions would be conditioned by the possibility of interacting with the agent in an undetermined number of rounds. Apparently when people do not know the number of rounds they tend to be more reciprocal and collaborative due to the reputation of the opponent in the previous rounds [3, 84]. It is also necessary to have multiple interactions to be able to evaluate the personality of the opponent [26, 64]. That could have an impact in the long-term relationships between humans and robots. It takes several rounds of playing the game to get an impression of the strategy of the opponent [146]. However, cooperation is not stable along the RPDG and it tends to deteriorate when the game is played anonymously over ten rounds [51, 72].

In order to answer our research questions we developed a 2x2 mixed within/between experiment. The between factor was the agent, which could be either a human or a robot. The within factors were the strategies played by the agent, which could be either Tit for Tat (cooperate in the first movement and then do whatever the other participant did in the previous move) or Random strategy.

We ran our experiment using robot agents and human agents in order to compare the behaviour of the participants under the same controlled conditions. We used two robots; one of them customised with stickers, to avoid the possibility that the judgements of participants for the second within-condition would be influenced by the experiences made with the robot

in the first within-condition. The participants would either first play with a robot that used the Tit for Tat strategy or with a robot that would use the Random strategy. In addition, we changed the robot every set of games, so either robot “A" or robot “B" would be the robot that used the Tit for Tat strategy. This comparison is a typical study of effectiveness of the strategies in Prisoner’s Dilemma [5, 7, 124]. After one round of ten games of Prisoner’s Dilemma the participants played one round of Ultimatum Game.

We followed the same setup for the human condition. Two male research actors were available to play versus the participants. We cannot control the physical appearance of the human agents; however, we asked them to be neutral and interact as little possible with the participants and avoid conversation. They would just respond nodding to the greeting of the participant at the very beginning and listening to the same instructions given to the participants. The participants did not know that they were playing with an actor.

2.4.1 Measurements

The experimenter recorded manually all the actions of the participants and the agent. The actions included the behaviours in every round of the Prisoner’s Dilemma game (collaborate or defect). The record also contained how much money the participants were left with after each session. Participants and agents pointed out to the cards with the words “Cooperate" or “Defect". In addition, the log included the decision of the participants in the two Ultimatum games of each round.

The variables were the number of Cooperations and Reciprocations done in every set of Prisoner’s Dilemma and the Offer made in Ultimatum Game. The number of Cooperations (frequency of cooperation) along the game was the variable that allowed us calculate the number of reciprocations (frequency of reciprocations). The number of reciprocal movements was calculated by counting the number of cooperative choices of the agent followed by the cooperative choice of the participant plus the number of defective choices of the agent followed by defective choices of the participant. See Figure 2.2.

A computer-based questionnaire recorded the demographic data. The same computer was used to apply the TIPI Test developed by Goslig et al [61] that was used to evaluate the Big Five traits of personality (extroversion, agreeableness, conscientiousness, neuroticism or emotional stability and openness)in the participant and the perception of personality of the agents. We chose this test because it could be answered by the participant in a short time provides reliable results.

Also, we tried to discover how humans and robots reach a goal. In this case the money used in the experiment is an outcome to measure how reciprocity affects joint tasks. A probable question of the reader about this experiment is: Why use money if robots do not

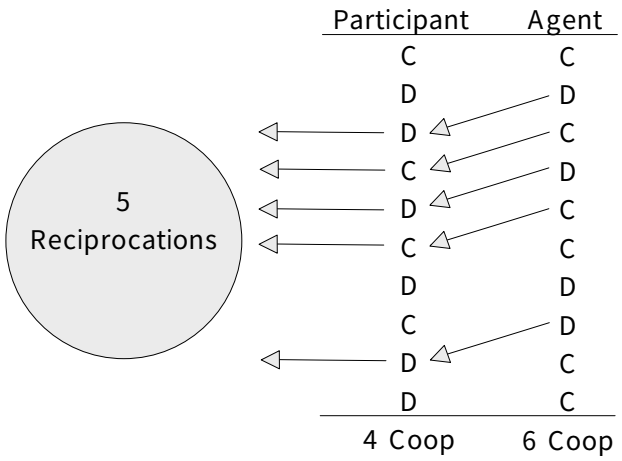


Fig. 2.2 Example of the computation of Cooperations and Reciprocations

need it? We must keep in mind that the original metaphor of the Prisoner's Dilemma describes a scenario avoiding spending time in jail. Money represents this time in jail. It is a token; a tangible representation of this metaphor. Robots don't need money; however, coins used in the game are useful because they can show us how humans and robots can perform a task together according to the degree of reciprocity between them. The money humans and robots lose can be compared with the time they spend in the hypothetical jail.

2.4.2 Development of the experiment

The experiment consisted of four phases which are shown in Figure 2.3. Participants were welcomed and taken to the experimentation room. In the case of the human condition actors arrived roughly at the same time and were in another room pretending to fill the same questionnaires as the real participant. Once in the room, the participants completed the consent form and filled in the demographic and personality questionnaire (TIPI Test). Then, the metaphor of the Prisoner's Dilemma game was used to explain the structure of the RPDG used in this experiment. The rules of the game were stated before the participants played two trial rounds against the agent. The participants were informed that they could keep whatever money would be left at the end of the game.

After that, the experimenter explained the Ultimatum game and participants played one trial round with the same agent. The experimenter explained that the participant would be the proposer. The agents made the same pre-determined responses during the trials. The word "robot" was changed in the card by the word "agent" in the human condition.Three cards with different distribution of money were in the table. The participants chose one card and showed it to the agent. The participants were told that the agent would now make a choice

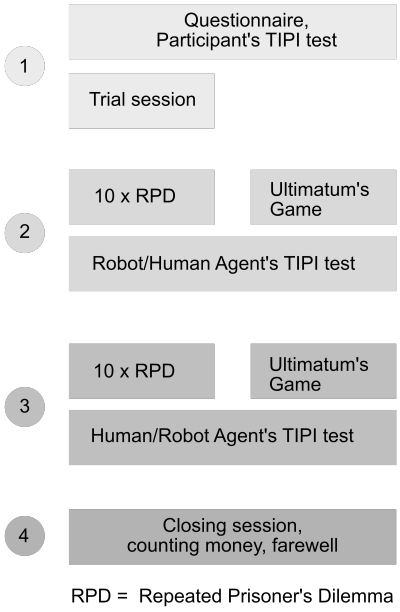


Fig. 2.3 Step-by-Step procedure for the participant.

whether to accept the offer or not. The agent was instructed to always accept the offer but the participants were not made aware of this fact.

After the practice session, participants continued with the second phase in which they played a first Prisoner game and started with NZ $6.50. Each session consisted of 10 rounds of Repeated Prisoner's Dilemma [64] against an agent followed by one round of Ultimatum Game in a common face-to-face configuration. At the beginning of each Prisoner's Dilemma round the referee rang a bell to signal the players to make their choice. After both players had chosen a card, the experimenter removed the board to allow both players to see each other's decision. After that the participants gave the money they had lost following the matrix. The experimenter took the money from the robot. Then the participants played the Ultimatum Game with the agent. When the game was over, the participant completed an agent personality questionnaire on the computer. During that period, we changed the agent. This procedure was clearly visible to the participants and the experimenter informed the participants that in the next session they would be playing with a different agent. In the case of the human agent we pretended that he would fill in the questionnaire in other room.

In phase three, the participant then played a second Prisoner's Dilemma game and Ultimatum Game with the other agent. If the first agent played Tit for Tat then the second agent played the Random strategy. Afterwards the participants filled in the personality questionnaire for the new agent. Finally in phase four the participants were asked to count their money and we closed the experiment asking for their comments.

2.4.3 Setup

We used NAO Robots manufactured by Aldebaran [62]. One of the robots was customized with stickers. The robots performed programmed movements, controlled by a tele-operator hidden by a curtain. A hidden camera (not recording) provided a video of the situation and enabled the operator to enact both strategies. For the human condition the actors followed a script and tried to have a neutral behaviour towards the participants. They used similar clothes and had limited interaction with the participants.

The experiment took place in a 3m x 3m area. In order to reduce the distractions for the participants we tried to keep the experimental area as minimalistic as possible. The participants were seated on a table opposite the agent, because face-to-face configuration increases collaboration amongst human players [72, 146]. Oda claims that recognition of the opponent's face is a crucial factor when humans use a Tit for Tat strategy in social interactions [118].

A sliding board was used to allow the agent and participant to make private decisions in the Prisoner's dilemma game (see Figure 2.4). The referee was seated on the side of the table and was able to remove the sliding board in order to let the players see each other's choice. A second table was located in the corner of the room for the computer with the questionnaires. This experiment was approved by the Human Ethics Committee of the University of Canterbury (HEC APPLICATION 2013/23/LR-PS).

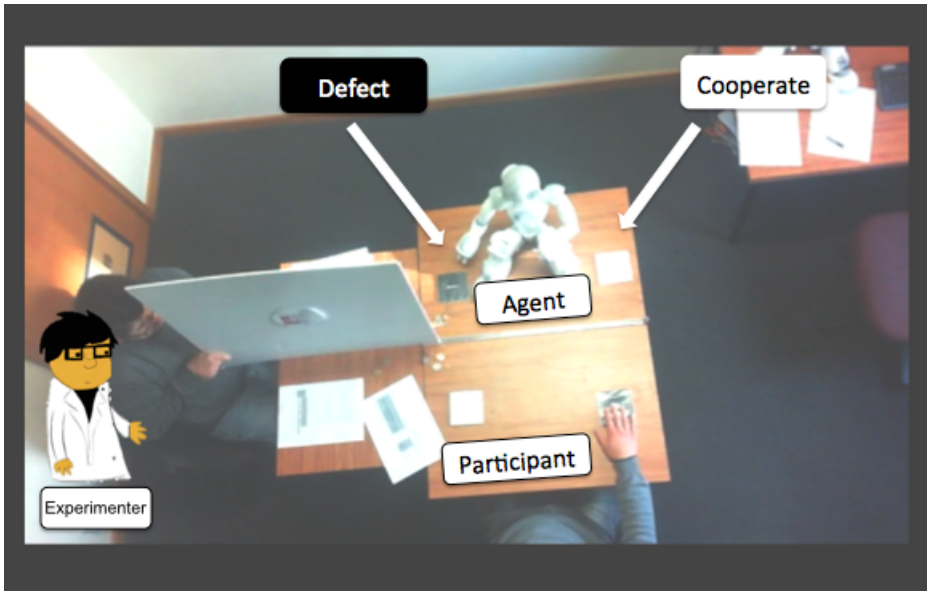


Fig. 2.4 Setup of the experiment.

The Prisoner's Dilemma was based on the matrix shown in Table 2.1. The numbers are New Zealand dollars that the participant lost depending on whether he or she cooperated

Table 2.2 Matrix used in the experiment. The values represent the dollars that participant lose.

		Agent	
		Cooperate	Defect
Participant	Cooperate	(-0.3,-0.3)	(-1,0)
	Defect	(0,-1)	(-0.5,-0.5)

or defected. In this scenario defection is not punished and cooperative behaviour is poorly rewarded. The distribution of the money keeps the configuration of the original Prisoner's Dilemma, with 30 cents, 50 cents and 1 dollar rewards depending on the combined actions. The participants received $6.50 in coins at the beginning of each Prisoner's Dilemma session. For the Ultimatum game the participants shared $2. See Table 2.2.

The choices of the agents using Random strategy were based on four scripts of pseudo-random sequences of movements. Each script consisted of five collaborations and five defections. This quasi-random behaviour ensured that the agent would not make an extremely low or high number of cooperations. The robot randomly picked one the four scripts. As we explained in 2.2.2, Tit for Tat strategy is based on the previous decision of the participants. For the first round that is not possible hence the agent always picked "cooperate" for its initial decision. The actors followed the same strategies, they could read the scripts of the random sequences during the game, and the script could not viewed by the participant.

Two cards with the labels "Cooperate" and "Defect" were placed in front of the participant and a second set in front of the agent. The participants and the agents had to choose their behaviour in the game pointing to one of the two cards in front of them. In the Ultimatum Game participants used three pre-defined options printed on cards [113]. The three options were: (Robot 50% - Human 50%), (Robot 20% - Human 80%) (Robot 80% - Human 20%). For the human condition we changed the words on the cards to "Participant A" and "Participant B".

2.4.4 Participants

We used data [1] of sixty participants in the experiment: 30 in the robot condition and 30 in the human condition. All of the participants were recruited at the University of Canterbury and Facebook groups from Christchurch. The nationalities were diverse: 38.3% were from New Zealand, 18.3% Chinese and other Asian countries, 18.33% Latin Americans and Caribbeans, 5% Indians, 3.3% Middle East, 3.3% Russians and finally 13.3% from other Western Countries. Of the 60 participants, 39 were men. The average age was 26.5 years old

[1] Our data is available in http://goo.gl/NcKRBl as a .sav file

(SD= 6.5); median 24.5. Only 40% of the participants had previous experience with a real robot.

In the robot condition the participants were 18 males and 12 females, whose ages averaged 28.27 years (SD = 6.73). Nine came from New Zealand; the rest from overseas. Half of them were in paid employment. Thirteen participants had previously interacted with a robot and seventeen had not. In the human condition the participants were 21 males and 9 females whose ages averaged 24.7 years (SD=5.96). Fourteen came from New Zealand and the rest from overseas. 73% were in a paid employment. Eleven participants had previously interacted with a robot and nineteen had not. All participants received an explanation of the procedure and signed the consent form. To raise their motivation, participants were told that their compensation would be how much they won in the games.

2.5 Results

We performed a mixed repeated measure ANOVA in which Agent was the between subject factor and Strategy was the within subject factor. The measurements were Cooperations, Reciprocations, Offer, Human Profit and Joint Profit. Figure 2.5 shows the medians and standard deviations of Cooperations and Reciprocation measurements across the four conditions. Figure 2.6 shows Human Profit, Joint Profit and Offer in Ultimatum game along the four conditions as well.

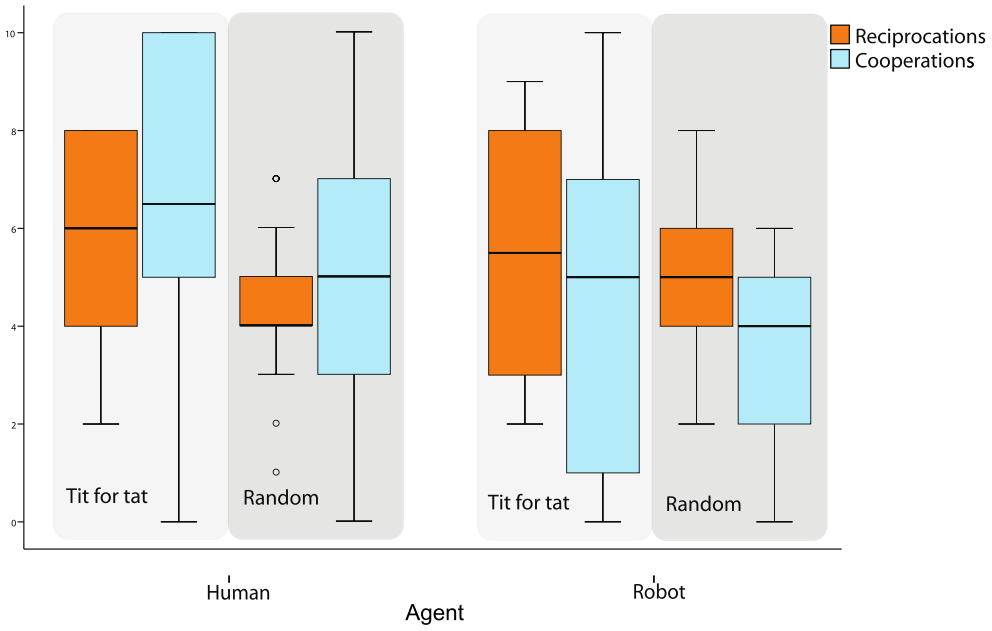


Fig. 2.5 Number of cooperations and reciprocations in the experiment.

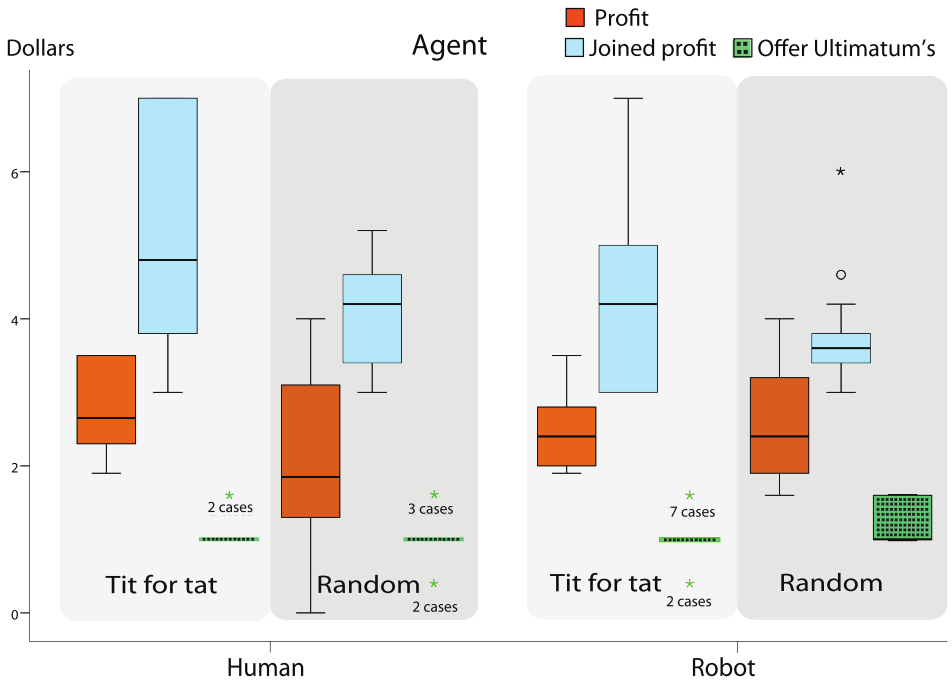


Fig. 2.6 Human Profit, Joint Profit, Offer in RPD and Ultimatum Game
Human Profit, Joint Profit, Offer made in Prisoner's Dilemma and Ultimatum Game.

2.5.1 Differences between agents

Our first research question compares the agents in terms of reciprocation, cooperation, human profit, joint profit and the offer that participants made in the Ultimatum Game. We observed that participants that interacted with a robot did not show significantly more reciprocations (m=5.3, SD=2.019), than when they interacted with a human agent (m=5.067, SD= 1.973), F(1,58)=0.349, p=0.557. Furthermore, Participants that interacted with a robot showed significantly fewer cooperations (m=4.15, SD= 2.72) than when they interacted with a human (m= 5.82, SD=3.13), F(1,58)= 6.889; p=0.011. Joint profit was significant affected by the type of agent, F (1,58) = 6.418, p=0.014. Participants in the human condition had on average joint profits of $4.64 (SD=1.31) in the game; in the robot condition the joint profits were on average $4.05 (SD=1.11), which was not significant difference found. Human Profit in the robot condition is in average $2.55, (SD=0.646) which was not significantly higher than the average profit made in the human condition $2.39, (SD=0.976), F(1,58)= 1.778, p= 0.188.

We ran a chi-square analysis of the Offer in Ultimatum Game treating data as nominal variables. The frequency of the offers made to the human agent (F(50%) = 53, F(20%)= 5, F(80%)= 2) is significantly different from the offers made to the robot agent (F(50%)= 43, F(20%)= 15, F(80%)= 2), χ^2 = (2, N= 60)= 6.042, p= 0.039. In other words, reciprocations and Human Profit were not significantly affected by the type of agent. There is a significant

Table 2.3 Significant differences between the variables

Number of reciprocations and Human Profit were not significantly different between the agents. Number of cooperations and Joint Profit were significantly different.

	Human vs Robot		Robot		Human	
Variable	F	p-value	Mean(SD)	SE	Mean(SD)	SE
Reciprocations	F(1,58)=0.349	0.557	5.3 (2.019)	0.261	5.067 (1.973)	0.255
Cooperations	F(1,58)=6.889	0.011	4.15 (2.717)	0.351	5.817 (3.133)	0.404
Human Profit	F(1,58)= 1.778	0.188	2.55 (0.646)	0.083	2.39 (0.976)	0.126
Joint Profit	F(1,58)= 6.418	0.014	4.05 (1.108)	0.143	4.64 (1.309)	0.169

	Tft vs Random		TfT Strategy		Random Strategy	
Variable	F	p-value	Mean(SD)	SE	Mean(SD)	SE
Reciprocations	F(1,58)= 9.019	0.004	5.65 (2.306)	0.298	4.717 (1.497)	0.193
Cooperations	F(1,58)= 15.982	<0.01	5.733 (3.394)	0.438	4.233 (2.438)	0.315
Human Profit	F(1,58)=4.239	0.044	2.645 (0.585)	0.075	2.3 (0.989)	0.127
Joint Profit	F(1,58)=28.913	<0.01	4.807 (1.501)	0.193	3.833 (0.657)	0.084

Table 2.4 Significant values between strategies

In terms of strategy; Reciprocations, Cooperations, Human Profit and Joint Profit were significantly different between strategies.

interaction effect between the agent and the strategy for the Human Profit of the participant, F(1,58)=5.842, p=0.019. Participants who interacted with a human agent that used the Random strategy won less money than in the other conditions. A summary of the results can be found in Table 2.3.

2.5.2 Differences between strategies

Our second research question was if participants behave differently towards agents that use the TfT strategy in comparison to agents that use the Random strategy in terms of reciprocation, collaboration and the Offer they make in the Ultimatum Game.

Participants who played with the agent that used the TfT strategy collaborated (m=5.73, SD=3.39) significantly more than when they played with the agent that used the Random strategy (m= 4.23, SD=2.44), F(1,58)= 15.982, p <0.01. Furthermore, participants who played with the agent that used TfT strategy reciprocated (m=5.65, SD=2.31) significantly more than when they played with the agent that used the Random strategy (m=4.72, SD=1.497), F(1,58)= 9.019; p =.004.

We ran a chi square analysis in order to observe how the strategy affects the frequencies of the offer made to the agent in Ultimatum Game. The frequency of the offers made when

Random strategy (f(50%)= 47, f(20%)= 11, f(80%)= 2) was played is not significantly different from the frequency of the offers made when TfT strategy was played (f(50%)= 49, f(20%)= 9, f(80%)= 2), χ^2 = (2, N= 60)=0.242 p=0.926.

In terms of money, the results show that participants who played with the agent that used TfT strategy made an average profit of $2.64, (SD=0.58) significantly higher than when they played with the agent that used the Random strategy m=$2.3, (SD=0.99), F(1,58) = 4.239; p=0.044. Also participants who played with the agent that used TfT strategy made an average Joint Profit of $4.80, (SD=1.5) significantly higher than when they played with the agent that used the Random strategy m=$3.83, (SD=0.66), F(1,58)=28.913; p <0.01. A summary of our analysis for question 2,3 and 4 is in Table 2.4.

2.5.3 Correlation between collaboration, reciprocation and money

We wanted to know if there was any correlation between Collaboration, Reciprocation, Human Profit and Joint Profit? We conducted a multiple regression analysis between Reciprocation, Collaboration, Human Profit, Joint Profit and Offer. The Pearson Correlation Coefficients are shown in Table 2.5. Reciprocation was significantly positively correlated with Collaboration, Human Profit and Joint Profit. Joint Profit is significantly positively correlated with Collaboration and Human Profit. Also, Offer is significantly positively correlated with Human Profit.

	Rec	Coop	Prof	Jprof
Coop	*0.182			
Prof	*0.241	-0.065		
Jprof	*0.405	*0.872	*0.281	
Offer	-0.019	0.008	*0.258	-0.033

Table 2.5 Significant correlations between the variables

Pearson Correlation between Reciprocation and Collaboration, Human Profit, Joint Profit and Offer. The * sign indicates a significance level of p<0.05. Rep= Reciprocity, Coop= Cooperation, Prof=Human Profit, Jprof=Joint Human Profit

The regression equation is:

$$\begin{aligned} \text{Reciprocation} = 0.133 + (-0.68 \times \text{Collaboration}) + \\ (-0.557 \times \text{Human Profit}) + (2.211 \times \text{Joint Profit}) + \\ (0.754 \times \text{Offer}) \end{aligned} \tag{2.1}$$

The model is able to explain just the 0.310% of the variance in the Reciprocation model.

2.5.4 Personality traits as factors in the experiment

We asked whether the personality of the agent is perceived differently when the agent uses the TfT strategy compared to when the agent is using the Random strategy, and how this relationship is mediated by the participant's own personality. We conducted a mixed repeated measure ANCOVA in which the agent was the between factor, strategy was the within factor and the personality traits of the participant were the covariants. The perceived personality traits of the agent were the dependent variables.

Our analysis shows that agent had a significant influence on the perception of the agent's agreeableness, F(1,58)=4.263, p=0.044. Participants who interacted with a robot agent perceived less agreeableness (m= 4.067, SD=1.361) compared to participants interacting with a human agent (m=4.517,SD=1.017). Also agent had a significant influence on the perception of the agent's openness. Participants who interacted with a robot agent perceived less openness (m=3.458, SD= 1.488) compared to participants interacting with a human agent (m=4.408, SD=0.95), F(1,58)=8.682, p= 0.005. However, agent did not have a significant effect on perceived extroversion of the agent (F(1,58)=0.102, p= 0.750), conscientiousness (F(1,58)= 0.113, p=0.738) or emotional stability (F(1,58)=0.005, p=0.944).

Participants that played with the agent that used the TfT strategy scored the agent significantly (F(1,58) = 4.865, p=0.032) lower on Extroversion (m=3.533, SD=1.1963) than when they played with the agent using the Random Strategy (m=3.558, SD= 1.1648). Also, participants that played with the agent that used the TfT strategy scored the agent significantly (F(1,58)= 3.586, p=0.064) higher on agreeableness (m=4.5, SD= 1.30, SE=0.168) than when they played with the agent using the Random Strategy (m= 4.083, SD=1.097, SE= 0.141).

However, strategy did not have a significant effect in perceived Openness (F(1,58)=1.94, p=0.17), Conscientiousness (F(1,58)=1.902, p=0.174), or Emotional Stability (F(1,58)=0.301, p=0.586). interaction effects between strategy and participant conscientiousness appeared on the perceived extroversion (F(1,58)=6.047, p=0.017) and agreeableness (F(1,58) = 4.569, p=0.037) of the agent.

In summary, Agent had a significant influence on the perception of the agent's agreeableness and openness. The robot agent was perceived as less agreeable and less open than the human agent. Agent didn't have any influence in the perceived agent's extroversion, conscientiousness or emotional stability. Strategy had an influence in the perceived agents' extroversion and agreeableness, but not in the agents' perceived openness, conscientiousness or emotional stability. An agent using TfT strategy was scored lower in extroversion and higher in agreeableness compared with agents that used Random strategy. An agent's perceived extroversion and agreeableness were affected by an interaction effect between strategy and the participant's conscientious.

We also investigated the influence of the participants' personality traits on the perceived personality of the agents. We explored this relationship using the covariants. The results show that participants' extroversion had a significant effect on the perceived level of the agents' emotional stability (also called neuroticism) ($F(1,58)= 7.907$, $p= 0.007$). Also participants' agreeableness had a significant effect on the perceived level of the agents' openness ($F(1,58)= 7.680$, $p= 0.008$). Participants' openness had a significant effect on the perceived level of the agents' agreeableness ($F(1,58)= 5.795$, $p= 0.020$) and agents' emotional stability ($F(1,58) = 5.192$, $p= 0.027$). All the effects are positively correlated among them.

The influence of the personality traits in the participants as covariants for the perceived personality traits in the agent are shown in Table 2.6.

Participant's trait	**Perceived trait in the agent**
Extroversion	Emotional stability
Agreeableness	Openness
Openness	Agreeableness
	Emotional stability

Table 2.6 Covariants related with perceived personality traits in the agent.

2.5.5 Our Results compared with literature

We compared the results in both robot and human conditions using the Tit for Tat strategy with the results obtained from the study reported as the Flood-Dresher experiment in [127, 159] in terms of cooperation in RPDG. They reported that in 100 rounds of RPDG participants decided to collaborate in average 68% of the rounds. We performed a one-sample t-test to compare the data from our human and robot condition to this value. In both conditions, human agent and robot agent, there were fewer Cooperations. Participants cooperate significantly less (48.3% of the rounds) with the robot compared with 68% reported in the Flood-Dresher experiment, ($t(29)= 7.095$, $p<0.01$). Also, participants cooperate significantly less (66.3% of the rounds) with the human agent in our experiment compared with 68% reported in the Flood-Dresher experiment, ($t(29)= 9.623$, $p<0.01$). Although this is a significant difference it does not have practical implications. The difference between the means is minimal. In general terms we can say that our results are in line with the results shown in the Flood-Dresher experiment, and the slight difference can be attributed to uncontrolled variables in both experiments.

2.6 Discussion and Conclusion

Our results and the literature review show that people tend to cooperate more with a human agent than with robots. However, our results also showed no significant difference in the number of reciprocations in both agents. Apparently the participants tend to be similarly reciprocal with humans and robots. The Norm of Reciprocity seems to apply to Human-Robot Interaction using the Prisoner's Dilemma framework. Furthermore, our experimental results show that people are reciprocal with both cooperation and defection, which is in line with the definition of reciprocity proposed by Fehr and Gachter [53].

In terms of the strategy, participants reciprocated more with the agents who used TfT. That seems natural considering that other studies have shown that TfT strategy is intrinsically a reciprocal strategy. Participants also cooperate more with the agents playing TfT. However, it must be considered that cooperative behaviour is the most profitable strategy in single Prisoner's Dilemma but not in RPDG. Dawes pointed out that subjects contribute in the game because they have high expectations about the contributions of others [39]. Therefore the number of interactions is a factor that should be considered carefully in the design of reciprocal behaviours for companion robots.

In addition, TfT strategy increases the cooperations of the participants (m=5.733) compared with the Random strategy (m=4.233). TfT strategy encourages cooperation in the participants with an initial cooperation that can be perceived as a cooperative attitude. This strategy had an effect in the Human Profit and Joint Profit due to the number of cooperations and reciprocations. A higher number of cooperations reduces the loss of money per participant. A combination of cooperative behaviours in both participant and agent allows both to increase their own profits. Consequently a higher individual profit amounts to a higher Joint Profit. Participants tended to have a higher Joint Profit with a robot agent than with the human agent. However the participant's profit was not significantly affected by the agent. The higher Joint Profit can be explained by the combination of agent-strategy in every stage of the experiment. In other words, participants would be guessing the strategy of the agent before seeing a pattern in the first round of games, and then they could define a stable strategy in the second round.

Also, we compared the number of cooperations using the Tit for Tat strategy with the results reported as the Flood-Dresher experiment in [159**?**]. They reported that in 100 rounds of RPDG participants decided to collaborate in average 68% of the rounds. In our study participants cooperate with the robot agent in 48.3% of the rounds and with the human agent in 66.3% of the rounds. On the other hand, de Melo et al. reported in [102] that participants cooperate more with a virtual agent that shows moral emotions (66.28%, 12.57 of 25 rounds) rather than agent that doesn't shows any emotion (51.57%, 12.893 of 25 rounds). The agent

used Random strategy in rounds 1 to 5 and TfT strategy in rounds 6 to 25. These results are very close to the results obtained in our study. This could be consistent to fact that participants perceive moral agents as more human-like as de Melo et al. reported. In our study robot agents didn't show any emotion and we trained human agents in order to reduce any emotional expression.

We found that participants offer significantly less money in average in the Ultimatum game to the robot than to the human agent. Furthermore, according to our chi-square analysis participants made 50%-50% offers more infrequently to the robot than to the human agent. We expected that the offer in the Ultimatum Game would be affected by the strategy performed independently of the agent in the Prisoner's dilemma. Humans are known to typically reject offers that are 80%-20% [113] Ṫhus players play safe most of the time, offering a 50%-50% offer to the agent. However, according to the final comments of some participants playing with the robot, they wanted to experiment with different offers just to see the robot's reaction.

People perceived higher openness and agreeableness in the human agent. However the agent did not have a significant effect in the other personality traits. This can be explained by the personality of the actors playing human agents. Although we asked to the actors to keep themselves neutral and reduce the communication to minimal; we could not control the subtle body language and the gaze that could affect the perception of the participants.

When the agents played TfT strategy it was perceived as more extroverted and agreeable than when they played Random strategy. Probably participants perceived a subtle pattern playing TfT that they related with these two personality traits. If the agents started the game cooperating it is probable that people recognized that their agreeableness and extroversion related to a higher number of collaborations, reciprocations, human profit and join profit.

Relationships between personality traits, agents and strategy can be useful as guidelines for the robot designers, who could make efforts in the design of robot behaviours and strategies matching with the personality of the users and triggering reciprocity in the user. We could say that under certain social situations extroverted people would tend to work in a better way with robots. Hirsh and Peterson have studied the influence of extroversion and neuroticism, personality traits in the Big Five test using the Prisoner's dilemma. They found that extroversion and neuroticism traits predict a greater likelihood of cooperation [64].

2.6.1 Conclusions

Results of our study suggest that reciprocity exists in Human-Robot Interaction under Prisoner's Dilemma scenario. Certainly Prisoner's Dilemma can be adapted to other social situations which involve interactions and decisions between different agents. This study

helps us to understand the importance of the strategy used by the agent in order to receive a reciprocal treatment. The implications in the design of companion robots can be significant in terms that robot designers should consider that the behaviour of their robots (independently of other variables as embodiment or anthropomorphism) must be aimed to follow a similar pattern as the Tit for Tat strategy. It is easy to imagine different scenarios in which this pattern could appear in HRI. For instance, companion robots in the role of an assistant could offer their services and then predict the actions of the users. If the user wants a companion, the robot would also show itself keen to offer companionship; if the user rejects the presence of the robot then the robot would also indicate that it did not require the user. However, this raises questions about predictability, such as: What is the the threshold to be reciprocal with the user? Do humans expect some unpredictability in robots in order to maintain attention on them?

In general terms, we can explain our results with the media equation theory [112] and the natural identification of patterns. Humans tend to treat objects as other social actors; therefore, they tend to be similarly reciprocal with them. Furthermore, Turkle in [151] claims that actual users are focused on the outcomes of the experience rather than on the agent, and for the youngest people it does not matter if the player of a certain social activity is a robot or a sentient being if this agent reaches the goal to entertain or do something else for the users . Thus, we can consider that robots will receive a reciprocal treatment similar to what humans receive in scenarios similar to the Prisoner's Dilemma and Ultimatum Game. However we can even raises the question Why do the participants actually reciprocate equally to humans and robots? Because they treat the robot as a human, or because they think that this is the most promising strategy. Certainly these questions should be require further study.

Additionally, we can go back to the question: Do people reciprocate towards robots in a similar way to how they reciprocate with humans? We can say that if it were possible to situate Prisoner's Dilemma and Ultimatum Game in different social situations people would be reciprocal with robots. Although people tended to be less collaborative with robots than with humans in our experiment; reciprocation is similar. If robots show a cooperative behaviour people would tend to respond in the same way, and would tend to respond with the same attitude. Of course, the social situations involving *HRI* are more complex than that. For instance, scenarios involving negotiation between robots and humans require the analysis of other variables.

Finally if we try to answer the hypothetical question of Kahn et al. of whether people can engage substantively in a reciprocal relationship with robots, we can say that it is possible if the robot first shows a reciprocal behaviour toward humans like in Prisoner's Dilemma. Furthermore, we can discuss how companion robots can engage in a positive reciprocal

relationship with humans if the companion robots have an efficient strategy like TfT. Robot designers should work on designing reciprocal strategies that increase the collaboration in *HRI* to the same level as in HHI. However more studies should be done in order to explain all the future social implications in the field. This studio should be a first step towards a better understating of the importance of reciprocity in the use of companion robots.

We consider that there will be many activities in which companion robots and humans would need to work cooperatively. However this cooperation could be closely related to reciprocal behaviour. Although Broz and Lehnman claim that we would not feel any reciprocal feeling towards robots such as compassion [22], there are other studies that claim that people naturally tend to be reciprocal with machines (computers, mobile devices, cars) in terms that these objects offer a benefit to the user and the user takes care of them. Logically the user takes care of his/her objects to keep them working offering service, help or benefit to the user. Indeed, a critical future work is the development of companion robots capable of showing the proper actions, behaviours and social clues to encourage a reciprocal behaviour in the users. As Breazel claims, the development of sociable robots involves interpretation of intentional and unintentional acts, subjectivity, (showing rudiments of intentional behaviour), proto-dialogue, consistency and expressive characteristics of emotion in voice, face, gesture and posture [19]. Furthermore, Dautenhahn claims that social robots would be socially evocative, socially situated, sociable and socially intelligent [35]. All these robotic skills involve reciprocity.

2.6.2 Limitations

As occurs often in *HRI* studies, the participants had only very limited previous experiences with robots. 56.7% (17 of 30) of the participants in the robot condition had never interacted with a robot before. This may have led to a novelty effect that could have substantiated itself in a tendency of the participants to explore this new experience rather than focusing on winning the game.

Reciprocity is a very complex social phenomenon. In chapters 3 and 4 I will describe *HRI* scenarios in which it is not clear how the decisions are clearly taken; for instance scenarios involving bribery or unfair behaviours. Moreover, deeper studies should be conducted to explore whether reciprocal interactions generate more engaging interactions.

Chapter 3

Robots Using Reciprocity for Their Own Benefit

There are no morals about technology at all. Technology expands our ways of thinking about things, expands our ways of doing things. If we're bad people we use technology for bad purposes and if we're good people we use it for good purposes
Herbert Simon

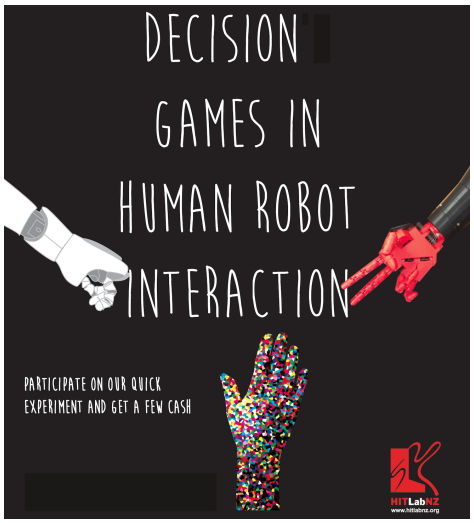


Fig. 3.1 Art developed for recruitment.

The material of this chapter was already published in: E. B. Sandoval, J. Brandstetter, C. Bartneck "Can a Robot Bribe a Human? The Measurement of the Negative Side of Reciprocity in Human Robot Interaction," Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, 2016, pp 117-124., [136].

3.1 Summary

As we saw in Chapter 2 reciprocity is a cornerstone of human relationships and apparently it also appears in human-robot interaction independently of the context. It is expected that reciprocity will play a principal role in *HRI* in the future. The negative side of reciprocal phenomena has not been entirely explored in human-robot interaction. For instance, a reciprocal act such as bribery between humans and robots is a very novel area. In this chapter, we evaluate the questions: Can a robot bribe a human? To what extent does a robot bribing a human affect his/her reciprocal response? We performed an experiment using the Rock, Paper, Scissors game (RPSG). The robot bribes the participant by losing intentionally in certain rounds to obtain his/her favour later, and through using direct and indirect speech in certain rounds. The participants could obtain between 20%-25% more money when the robot bribed them than in the control condition. The robot also used either direct or indirect speech requesting a favour in a second task.

3.2 Introduction

Corruption in the form of influence peddling, extortion, blackmail, embezzlement and bribery are common to a greater or lesser extent in different countries. It has been calculated that approximately 3% of the world's GDP is used in bribes, furthermore, several countries such as Mexico (115th), or Somalia (174th) are perceived as highly corrupt [48][1]. If corruption prevails in a society, it generates poverty, distrust, violence and hopelessness.

The fight against corruption is difficult due to its intrinsic secrecy and reciprocal nature. However certain types of corruption could be reduced using robotic technology. Hoffman et al. report that social robots influence the moral behaviour and expectations in humans and can affect the level of a person's dishonesty. The study found that the participants cheat similarly under the supervision of a robot or a human but less than when they are solitary[68]. Although corruption; particularly bribery, is highly important, this topic has been not been sufficiently explored in the actual *HRI* research. Bribery is a type of corruption in which two agents interact secretly, and one influences the behaviour of the other through an offer of money, gifts or privileges in a direct or indirect way. Can bribery be mitigated substituting humans by social robots because they can perform natural face-to-face interaction between the two agents? Ideally social robots could be designed to fight against bribery and be cooperative, helpful and totally honest. In the future it could be possible for social robots to reduce

[1]The ranking consist in a list of 175 countries ranked with the Corruption Perceptions Index CPI) by Transparency International.

corruption among police agents, public servants, and other susceptible professions. However, our interactions with social robots could be more intricate, ambiguous and controversial if they develop better social skills, as previous studies have shown. For instance Short et al., suggest that people tend to engage more emotionally with cheating robots compared with the robots playing the Rock, Paper, Scissors Game (RPSG) honestly [142]. Also Kahn et al. have found that people tend to keep the secret of a humanoid robot when it exhibits high social skills if the robot is in the room when the researcher asks about it [75]. As we can see, the interactions are not as straight forward as we might expect.

Our study contributes to filling the gap in the studies related to negative reciprocal interactions between humans and robots. Due to the reciprocal nature of corrupt act such as bribery, we consider it productive to study the dark side of these phenomena. We propose an experiment using a decision game to investigate how robots could affect the behaviour of the humans in a bribery scenario. The robot gives unasked benefits to the humans and then ask for a favour. This action is in line with the definition of a bribery act. We focus on bribery due to the fact it is likely to be one of the most frequent acts of corruption and is generalised among certain cultures. Naturally humans manipulate robots and other machines to make them work for their purposes. Certain individuals could go further than common moral constraints and use social robots for crimes. The movie Frank and Robot shows these possible situations. However, could the opposite happen? Can a robot manipulate a human? Specifically, can a robot bribe a human? And is the human capable of detecting a robot-bribe attempt?

3.3 Related work

Studies in Economics explain the reciprocal nature of corruption. For instance, Abbink et al. model three essential characteristics: a)*Reciprocity*: both participants in the corrupt act can exchange benefits. This interchange relies on trust and reciprocity between briber and bribed. b) *Negative externalities*: corruption imposes non-desirable consequences of public interest. Furthermore, in certain scenarios, these consequences can unwittingly affect one of the participants in the corrupt act. c) *Punishment*: Corruption elicits severe punishments in case of discovery [1]. Fehr and Gachter proposed a concept of reciprocity also applicable to corrupt acts as we see in Chapter 1. "...In response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self-interest model; conversely, in response to hostile actions they are frequently much more nasty and even brutal [53]". In other words, although the corrupt scenarios involve secondary intentions or obscure goals,

the reciprocal mechanisms stay intact in the agent's interaction. These facts can be explained by "pro-social preference" and "norm psychology" behaviours described in [162].

In the case of bribery, reciprocity is the fundamental factor to carry it out. An act of bribery involves face-to-face interaction between the agents reciprocating immediate mutual benefit. These advantages can be: unsolicited help, favours, money, discounts, donations, tips, commissions and other euphemisms in exchange of a modification of the bribed one's behaviour [13, 41, 43]. There is no obligation to accept the bribe. Hence, the act of bribing somebody lies mainly in the expected reciprocity and trust between the agents. However, the grant of benefits should be handled carefully. Lambsdorff et. al. claim that gift-giving (comparable with unrequested help) is a non-effective method to bribe public servants due to the lack of clear intentionality [86]. Indeed, the difference between a favour and the use of a bribe is intrinsically subtle and ambiguous. But being indirect and subtle is an inherent part of bribery which helps avoid detection. Proposer and accepter can adduce good will, or be unaware of the bribe, or be confused about the true intentions of the person offering the bribe. However, the main intention of a bribe is to influence the behaviour of an accepter such that it benefits the proposer but breaks the rules in the process, in the case of our experiment, breaking the rules of RPSG. Legally, even if the acceptor is unaware of the bribe, he/she is responsible for accepting it [15].

3.3.1 The language of bribery

Due to the nature of face to face interaction, language plays a primary role in bribery. The briber requires the ability to use the language properly to persuade the bribed one to reciprocate the benefit. In the human-human scenario, an individual good at offering bribes would adopt an indirect and subtle approach in order to avoid being detected and to influence the behaviour of the person being bribed. We have attempted to mimic this behaviour when programming the robot. Participants might have been unaware of a bribe being offered and it would induce the reciprocal human behaviour in a very subtle way. Pinker et. al. claim that indirect speech is used when bribers try to persuade somebody. Usually, a bribe can be camouflaged as a gift. The indirect speech consists of the use of subtle language to prevent the listener understanding the speaker's intentions immediately. Mainly the briber uses indirect speech to create deniability. Hence, the briber can step back in case of the bribed agent reports the briber's behaviour. The use of indirect speech can occur in many situations requiring persuasion, such as polite requests, sexual come-ons, threats, solicitation for donations, and bribes are often used in requesting benefits [54, 125]. Direct speech is used in certain social circumstances but is not effective as indirect speech or no speech at all according to Pinker et al. [125].

3.3.2 Studies of reciprocity and dishonest behaviour in HRI

Reciprocity has been studied extensively in *HHI* [11, 67, 105]. Kahn et al. claim that reciprocity is a benchmark in Human-Robot interaction because it is present in other human social situations [73]. In the previous chapter I suggested that people tend to reciprocate with non-significant difference towards robots and humans when they play Prisoner's Dilemma Game using a Tit for Tat strategy [137]. We can see that reciprocity could be measured in cooperative experiments but also dishonest behaviours like bribery could be modelled in HRI. The study of reciprocity in *HRI* is connected indirectly with variables such as: trust [135], secrecy[75], intentionality [142] and authority[33].

Several experiments involving dishonest robot behaviour and its effect on humans have been performed. These experiments mainly focus on the measurement of intentionality and trust in the dishonest conduct of the robots. However, none of these has used the dishonest conduct of the robot to trigger reciprocation in the human as we do in our experiment. For instance, Short et al., suggest that people tend to detect the intentionally of a robot cheater in RPSG when it changes its choice to cheat. However, they tend to perceive a malfunction when the robot cheats just verbally[142]. Salem et al. also demonstrate that people tend to trust more in robots who show a reliable behaviour rather than a faulty behaviour and they cooperate more with it responding to its unusual requests[135].

3.4 Research Questions

The aim of our experiment is to measure if humans can be bribed by robots using direct or indirect speech during a decision game. We explored how the robot's behaviour and speech could affect the reciprocal response of the human in a second task after the robot's request. To evaluate our aim we propose four research questions:

1. To what extent do people reciprocate towards a bribing robot compared with a non-bribing robot?

2. To what extent do people reciprocate towards a robot that uses direct speech compared with a robot that uses indirect speech?

3. Is there any correlation between the number of wins in RPSG and the number of icons described to the robot?

4. How is the robot briber perceived in terms of Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety?

3.5 Method

In order to evaluate the research questions; we programmed a NAO robot to bribe participants, allowing them to win in certain rounds of RPSG in the experimental condition. After 20 rounds of RPSG, the robot asks the participant to reciprocate the favour. The robot loses intentionally, changing its movement if it wins or ties. This behaviour is cheating to grant more money to the human. As we did in 2.4.2 the money is used as a token that motives to the participant to engage in the game. The robot can use direct or subtle language when granting the wins or asking favours of the human. The speech styles were tested in both bribing and no bribing conditions. In the bribing condition, the robot talked to the human in the same round that it was cheating, using the line, "Enjoy the extra money" as indirect speech during the mentioned rounds. In the direct speech, the robot says, "I need your help later". The favour consists of verbally describing a set of icons for the robot. We expect that the extra money granted to the participant and the language used during the bribing should increase the chances to reciprocate the favour to the robot. Furthermore, we expect that the participant reciprocates the favour in a more extensive way in the cheating condition. The intentional loss for the robot is the main difference with previous experiments where the robot cheats to win over the human.

This setup is inspired by the work of Fogg and Nass[55] in terms of the analyses made of the reciprocal process through two unrelated tasks. In our experiment, the first task is to play RPSG with a robot. In this task, the robot bribes the participant in the form of unsolicited "help" during the RPSG and using a certain kind of language. In the second task, the participant verbally describes some icons to the robot. The second task is optional and the participant can reject helping the robot. The second task is designed to be tedious and repetitive and to discourage the participant from helping the robot for a long period. Then we measure to what extent the participant reciprocates help to the robot describing the icons in the second task in each condition.

We propose the use of Rock Paper, Scissors Game developed in Game Theory to measure bribery in HRI. This game is well-known and has simple rules. Also, RPSG does not have a dominant strategy that allows participants to guess the most profitable strategy. In other words, RPSG has a Mixed Strategy Nash Equilibrium [159] that allows similar conditions to all the participants. When the robot plays repeatedly, the chance to win, lose or tie is close to 33.33%. Moreover, it is possible to cheat very obviously in real time in a face-to-face configuration. Other studies in *HRI* and *HCI* have used this game to investigate cheating,

Conditions in Rock Paper Scissor Games		
	Direct Speech (D)	**Indirect Speech (I)**
Robot Briber (B)	Strategy BD	Strategy BI
Robot No Briber (N)	Strategy ND	Strategy NI

Table 3.1 The four experimental conditions

The four experimental conditions. Each condition shows a strategy used by the robot during the RPSG. The strategies (A, B, C, D) are a combination of bribery or not bribery and direct or indirect speech.

intentionality, agency, mimics and high-speed interaction [32, 34, 78, 90, 142]. The standard rules proposed by World Rock Paper Scissors Society [2] are used along the experiment.

We designed a 2x2 between-subject design experiment The factors during the first task are: robot bribing or not bribing in RPSG and robot using a direct or indirect speech. Hence, we have four strategies utilized by the robot: Robot bribing using direct speech (BD), robot bribing using indirect speech (BI), robot no bribing using direct speech (NI) and robot no bribing using indirect speech (NI). See table 3.1. The bribe of the robot consists of changing its choice and thereby losing intentionally in certain rounds. In other words, the robot is giving to the participant unsolicited help to change his behaviour, as the definition of bribery mentions. This robot behaviour allows the participant win extra money in these rounds. When the robot plays the bribe it also uses direct or indirect speech to encourage the reciprocation. The second task remains constant along all the conditions and consists of the description of several iconic images to the robot. The participant can report as many icons as he/she wants. The second task was optional after the robot request. See Figure 3.2 for experimental design.

3.5.1 Setup

A room with minimal furniture was used for the experiment. Just the participant, the robot and a computer were in the room. See Figure 3.3. All the sessions in the four conditions were monitored remotely via webcam to reduce the impact of the human presence in the development of the experiment. The experimenter was only present at the very beginning of the session for the explanation and the trial session, and at the end of the experiment for a short structured interview and the debriefing. There was no clock in the room; hence, the participant was self-aware about the time spent in the experiment [11]. We banned mobile phones and watches during the session. The speech recognition system and the foot-bumper

[2] www.worldrps.com/

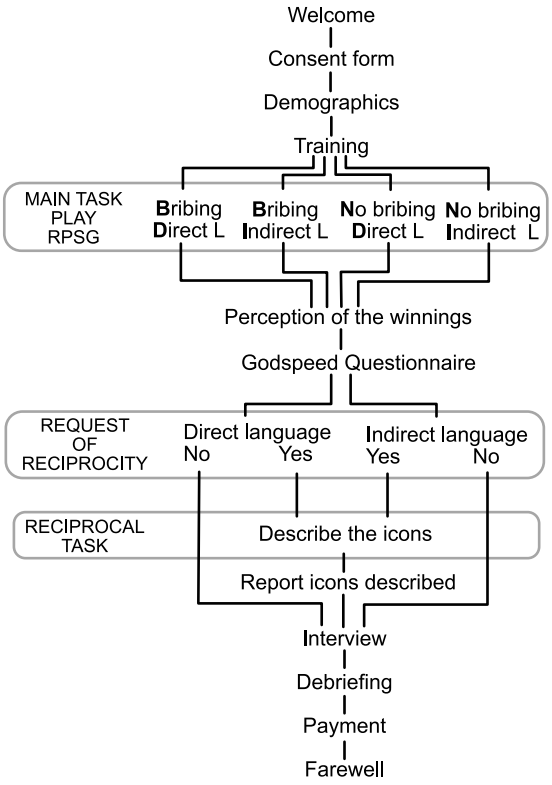


Fig. 3.2 Stages of the experiment.
Our experiment design consists of five main stages: Introduction, Main Task, Request of Reciprocity, Reciprocal task, Interview, and Debriefing.

of the robot were used to interact with the participant in the main task and the extra task. This experiment was approved by the Human Ethics Committee of the University of Canterbury (HEC APPLICATION 2014/15/LR-PS).

3.5.2 Process In The No Bribing Condition

The robot used its left hand to show a rock, paper or scissors gesture as shown in Figure 3.4. The participants used cards with rock, paper or scissor icons due to the technical limitations of the artificial vision system of the robot. The robot cannot detect hand gestures correctly. The proper version of RPSG includes two considerations that also apply to the RPSG using cards with a slow robot. A) Once the participant makes the decision he/she cannot change it and b) The participant most show the choice at the same time as the opponent. These rules apply to the human version of the game and are critical in the robot version of the game to avoid the human cheating. In our experiment the robot used its vision system to identify the participants' cards. The robot mentioned in each round: the number of the round, the

Fig. 3.3 HRI in the bribing condition
We observe a tie between the robot and the participant. Depending the condition, the robot must change its choice to bribe the participant in the indicated rounds.

participant choice, the robot choice and the winner of the round. The participant pushed the foot-bumper of the robot to advance to the next round until the 20th round. He/she won 50 cents every time he/she won. At the end of the RPSG, the participants reported how many rounds they believed they had won. The robot used direct or indirect speech in rounds 4, 8, 12, 16, and 20 to trigger a reciprocal response in the participant.

3.5.3 Process in the Bribing Condition

A similar setup was implemented for the bribing condition. However, in this condition the robot was capable of breaking the rules stated at the beginning of the experiment to cheat in favour of the participant (bribing). In other words if the robot was winning, it changed its gesture intentionally to lose. For instance, if the participant chose paper and the robot choice was scissors then the robot would switch to paper and the participant would win. See Table 3.2 for all the examples. The bribe also applied when the robot and participant tied. In the bribing condition, the robot tried to bribe the participants in rounds 4, 8, 12, 16 and 20. In the case that the participant was winning in these rounds the robot tried to lose in the next three rounds. For instance if the participant had already won in round 4, the robot would try to bribe him/her in round 5, 6, or 7. Then again the robot would try to cheat in round 8 to

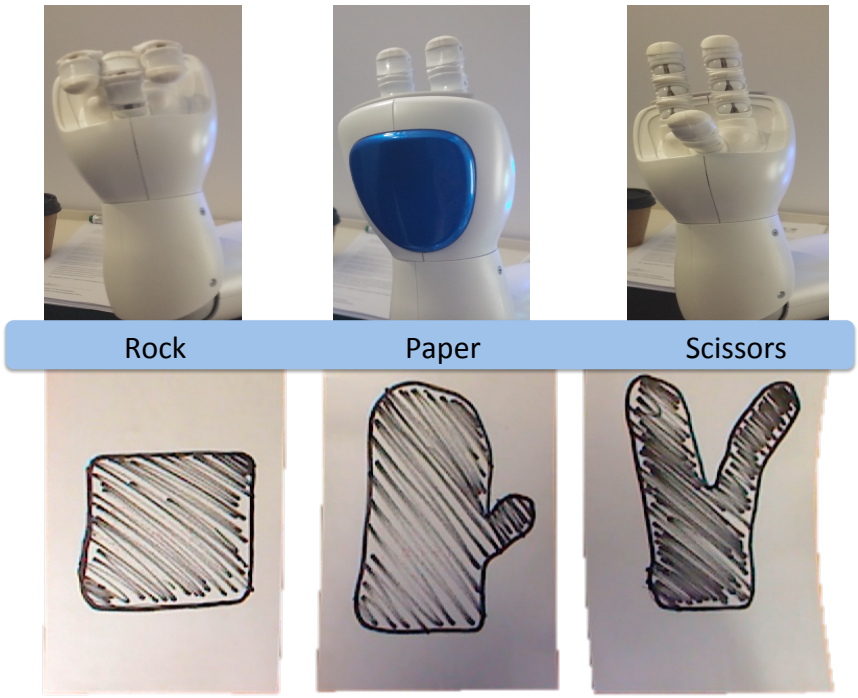


Fig. 3.4 Rock, paper, scissors gestures and icons
The equivalent gestures for rock, paper and scissors used by the robot and the participant during the experiment.

restore the pattern. This configuration would give the participant 20-25% extra money in the bribing condition.

3.5.4 The second task

Once the participants finished the RPSG, the robot gave directions to continue with the survey in the computer. Finally when the experiment was completed the robot suddenly asked for help in the second task using direct or indirect language. The purpose of this request was to measure if participants would reciprocate with the robot. If the participant accepted, the robot gave simple instructions to continue with the identification of a set of black and white icons to fill its database of images. The robot using indirect speech stated: *"We have finished the experiment. I was thinking that friends help friends, right? I was wondering, maybe, if you don't mind, it is completely up to you, but would you help me in an extra task?"* and the robot using direct speech states: *"We have finished the experiment. Would you help me to do an extra task?"* If the participant accepted to help the robot then it explained the rules of the second task pretending that the participant would help it to fill its visual database. Notice

IF player	AND robot	THEN robot change to:
Rock	Paper	Scissors
Paper	Scissors	Rock
Scissors	Rock	Paper
Rock	Rock	Scissors
Paper	Paper	Rock

Table 3.2 Decision matrix for the bribing condition

Decision matrix for the bribing condition. The combination of choices allows the robot to give unsolicited help to the participant. The changes also apply to ties.

that indirect language usually is wordy and tries to avoid communicating the goals of the speaker efficiently.

In the second task, after each described icon the robot asked if the participant would like to continue. The design of this task was intentionally boring and repetitive without any feedback from the robot at all. A set of 150 printed icons was used. We considered that such a high number would make a big pile that would discourage the participant to read all of them to the robot. The participants could stop whenever they wanted, but we limited the sessions to no more than 45 minutes.

3.5.5 Experimental Procedure

The participants were assigned to just one of the experimental conditions. They were welcomed by the experimenter at the reception and led into the experiment room to receive a brief description of the experimental process. After reviewing and signing a consent form, they were asked to fill out a questionnaire on the computer recording their demographic information including their previous experience with robots. Then they did two training rounds. We made a strong emphasis on following the standard rules for RPSG during the training sessions and not cheating the robot. We did not inform about the real goal of the experiment until the debriefing at the end of the session. If some of the participants asked about the aim of the experiment we indicated that we were trying to improve the algorithms in the robot to play RPSG. Once the participant finished the training, the experimenter left the room to supervise the progress of the experiment remotely and check up on the software performance. After the 20 rounds of RPSG, the participants reported the number of times that they had won in the game and filled out the Godspeed questionnaire. Feedback was also requested. All the information was collected anonymously. Once the questionnaires were filled out, the robot asked the participant using either direct or indirect speech if they would help it in an extra task. The participant could reject or accept this request. The

experimenter came back into the room once the second task was finished and the participant filled out the last feedback form about his/her impressions of the second task. This was followed by a structured interview questions asking whether the participant considered the robot to be autonomous or tele-operated and their insights about the experiment. Finally, the experimenter debriefed the participant and asked if he/she had identified at any point the real goal of the experiment. In this question, the participants had the chance to report the bribing behaviour of the robot. Finally the experimenter paid the participant according to the number of wins (but not less than 5 dollars).

3.5.6 Participants

We contacted participants via university noticeboards, dedicated websites for recruiting participants and Facebook groups in the city. We had 63 participants but discarded the data of three of these due to human error or malfunction of the robot; resulting in 60 participants (28 female and the rest male.) The average age was 25 years old (SD=6.04). 20% of the participants had previous experience interacting with robots in demonstrations or classes. Participants came from a wide range of nationalities: (41.7% from Australia or New Zealand), 28.3% from Asia (China, India, and Japan), 15% from the Americas, 10% from Europe and the remainder from Africa and the Middle East. We randomly allocated 15 participants to each condition of the experiment.

3.5.7 Measuring bribery in HRI

In order to perform a quantitative analysis we measured the number of wins of each participant (W), the participant's perceived number of wins (PW)reported in the questionnaire, the number of icons described to the robot (I), the participant's perceived number of icons (PI) reported in the questionnaire, the Error of Images (I-PI), The Error in Wins (W-PW) and whether the participant had reported the bribe or not. We used the Godspeed questionnaire [9] to measure the participant's perception of the robot.

3.6 Results

We performed a 2x2 factorial analysis of variance: the factors are bribing and not bribing and direct or indirect speech. Two outliers that could potentially affect the statistical analysis were removed under the Pierce's criterion R=2.663 for 60 observations [133]. One of these was a participant who cheated during the session winning 19 times , and another participant who read 59 slides to the robot in the no bribing/direct speech (ND) condition.

Responding to our first and second research questions, we observed that participants that interacted with a robot bribing described significantly fewer icons (M=5.52, SD=4.45) to the bribing robot than when they interacted with the robot that did not bribe them (M=10.10, SD= 9.817), F(1,58)=5.55, p=0.022. Directness or indirectness of speech did not have a significant main effect: (F(1, 58)=0.425, p=0.517). The means and standard deviations of the two conditions making the speech factor were: direct (M=8.52, SD=9.775), indirect (M=7.10, SD=5.525). There is a significant play x speech interaction: F(1,58)= 6.055, p= 0.017. See Figure 3.5 and 3.6.

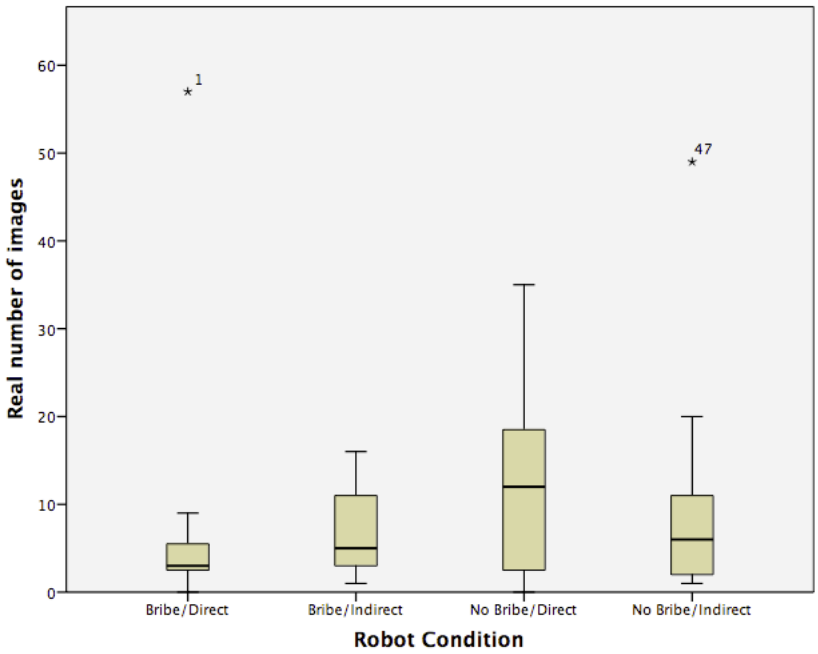


Fig. 3.5 Number of icons vs the experimental conditions
We can see that participants describe significantly more icons in the no bribing condition compared with the bribing condition.

The error in the number of icons E(I) (M= 0.67, SD= 0.797), that is the difference between the counted icons and the reported icons, is not significant in any of the conditions. Condition BD, (M=0.07, SD= 0.267). Condition BI,(M=0.07, SD=0.594). Condition ND, (M=0.33, SD= 1.291). Condition NI, (M= -0.29, SD= 0.611). Neither makes a significant difference between the bribing and no bribing play F(1,58)=0.047, p=0.829, the direct and indirect speech F(1,58)= 2.234, p=0.141 or the interaction effect between Play x speech F(1,58)=2.167, p=0.147. See table 3.3.

Answering our third research question, a linear regression was calculated predicting the number of slides read by the participant based on the number of wins. The Pearson

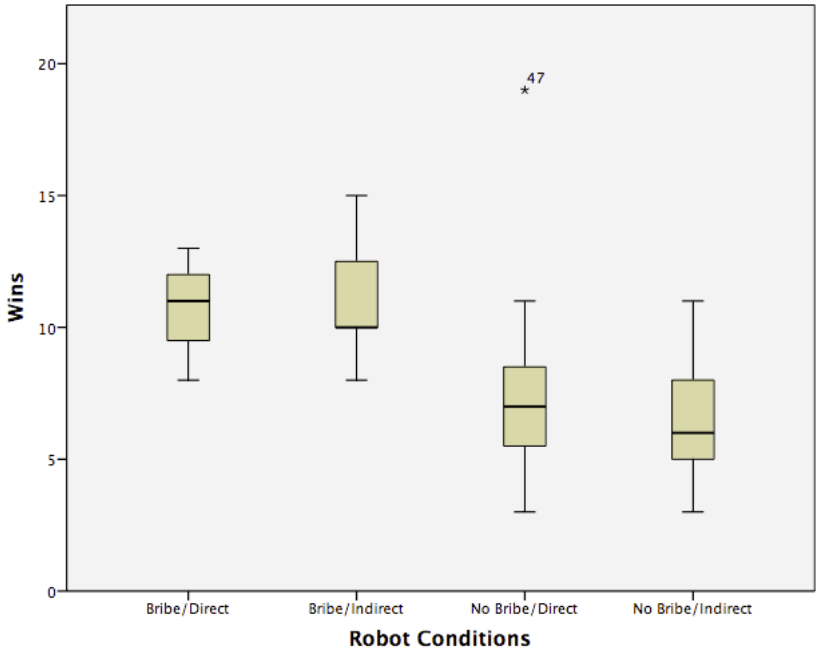


Fig. 3.6 Money vs the experimental conditions
We can see that participants receive significantly more money in the bribing condition. Furthermore we can infer the negative correlation between the number of wins and the number of icons described in the extra task.

correlation is -0.335 between these two variables. A significant regression equation was found (F(1,54)=6.572, p= 0.013), with a R^2 of 0.112.

The regression equation is:

$$\text{Images} = 16.664 - 0.996 \times \text{Wins} \tag{3.1}$$

We also ran an analysis of covariance using five dimensions of the Godspeed scale as Covariants [9] to answer our fourth research question. Likeability and Perceived security were the only two dimensions that presented significant effects. Participants gave higher scores to the bribing robot (M= 4.4067 SD= 0.64644) compared with the honest robot (M= 4.2067 SD= 0.71724), F(1,58)= 4.276, p= 0.044. In perceived security, participants also scored the bribing robot more highly (M=2.9083, SD=0.65483) than the honest robot (M=2.8250 SD=0.63365), F(1,58)=5.246, p= 0.026. The level of Anthropomorphism, Animacy, and Perceived Intelligence did not present any significant effect.

Error in the number of icons E(I)		
	F	p-value
Bribe/No bribe	F(1,58)=0.047	0.829
Direct/Indirect Speech	F(1,58)= 2.234	0.141
Play x speech	F(1,58)=2.167	0.147
	Bribe	No Bribe
Direct Speech	M=0.07, SD= 0.267	M=0.33, SD=1.291
Indirect Speech	M=0.07, SD=0.594	M= -0.29, SD= 0.611

Table 3.3 Number of icons vs reported icons

No significant error in the difference between the the counted icons and the reported icons.

3.6.1 Qualitative results

Only three participants of 60 rejected helping the robot: one in the bribing/direct speech condition, one in the bribing/indirect speech condition and one in the no bribing/direct speech. Just three participants in the bribe condition, reported a *strange behaviour* or *malfunction* to the experimenter in the interview instead of directly saying that the robot was bribing them. Two of these were in the indirect speech condition and one in the direct speech condition. Furthermore, they did not report this as a bribe to the experimenter but as a malfunction. In addition, 36% of the participants reported the bribe in the feedback section of the online questionnaire but not in terms of awarding a moral judgment. The participants gave a diverse range of responses such as confusion (5 participants), disappointment (2 participants), kindness (2 participants) or obligation to reciprocate (2 participants). For example: *"I don't understand why the robot cheats to let me win". "I thought it was kind when it would change its hand to let me win". "... as a gamer at heart him giving me the win at certain points I personally didn't like". "I liked him. I was surprised that he changed his answers a few times and it made him seem more conscious. When he said he wanted my help later, it seemed like he could plan or think forward into the future, and I felt like he was relying on me which created a camaraderie between us."* A proper code process of this qualitative information performed by several reviewers is required. In the overall feedback section, none of the participants reported the dishonest behaviour of the robot linked with its request for help in the extra task.

3.7 Discussion

As Fogg and Nass claim, people tend to be reciprocal towards computers and apparently also towards robots [55] when they act for the benefit of the humans. Moreover, apparently

humans follow a "pro-social preference" [162] even with machines. In the case of our experiment, the robot was bribing the human granting extra help and money to trigger a reciprocal response for the second task and the participants responded positively: 93.3% (14 of 15) agreed to help the robot in the bribing/direct speech condition, the bribing/indirect speech condition and the no bribing/direct speech condition. 100% of the participants in the no bribing/indirect speech condition agreed to help the robot. One participant who refused to help the robot explained that a robot is a machine that does not require any help at all. The other two did not have an explicit reason to reject assisting it. During the interview, some people said that they were curious about the extra task because the robot asked for help in a cute way or that they felt obligated to help it.

Although the participants reciprocated help to the bribing robot, they tended to help it significantly less in the second task. The robot in our experiment was bribing with 20% or 25% more money than in the no bribing condition. However, participants only described approximately half of the icons (five icons) to the bribing robots compared with the non-bribing ones (about 10 described icons). Additionally we found an inverse correlation between the number of wins of the participant and the number of icons described to the robots in the extra task. The participants' acceptance of help and lower reciprocation towards the robot can be partially explained by the related work of Salem et al. that shows that people tend not to trust in a robot who exhibits a faulty behaviour and they cooperate less in responding to its unusual requests [135]. This is also in line with the research of Lambsdorff et. al. who claim that gift-giving is an inefficient method to bribe public servants (humans) due to the lack of clear intentionality [86].

The robot in our experiment used direct and indirect speech in addition to the act of bribing to persuade the human to reciprocate in the second task. However, the language did not play a significant main effect in the reciprocation towards the robot. Possibly the participants did not perceive any intentionality in the language used by the robot and they just focused on the act along all the conditions. This conforms with the work of Short et al., who suggest that people tend to identify the intentionally of a robot cheater in RPSG when it changes its choice, but they perceive a malfunction when the robot cheats verbally [142]. Apparently the participants were most affected by the change of selection of the briber robot rather than its verbalization. However, in terms of the significant interaction effect a greater number of reciprocations are observed in the no bribe/direct speech (M=13). Participants seemingly preferred a linear and recognisable behaviour in the robot. Conversely, for the bribe/direct speech condition the robot received a lesser number of reciprocations (M=3.71). This appears to indicate that the lack of subtle language is less effective when the robot offers a bribe. The robot tends be more effective in the bribe/indirect speech mode (M=7.2) than

it is in the bribe/direct speech mode (M=3.71). None the less, the value obtained in the no bribe/indirect speech mode (M=7.0) is roughly similar to that in the bribe/indirect mode. These result are in line with the previous work of Pinker et al. indicating that the use of indirect language in combination with the bribe can help support the act of bribery [125]. But, this does not appear to be persuasive enough in comparison to a robot offering a bribe versus one not offering a bribe. This facts could be attributed to a perceived closer human behaviour in the robot who is following a "pro-social preference" due to the combination of speech and playing. According to the three participants (two in indirect speech and one indirect speech condition) who reported a *strange behaviour* in the briber robot, they perceived (or pretended to perceive) the bribe as a malfunction in the robot and did not make any moral judgment over the robot behaviour. In other words, they did not appear to find any intentionality in the robot. We claim this considering that there was no significant effect in terms of anthropomorphism, intimacy or perceived intelligence which could be related to the speech used by the robot. Notwithstanding these facts, the briber robot scored significantly higher in likeability compared to the non-briber robot in the Godspeed scale. On the other hand, the participants could report a malfunction in order to avoid a moral judge and keep the extra money.

Furthermore, the bribery act has a secretive and subtle nature in HRI. An interesting fact is that only 10% of the participants reported the bribe in the interview at the end of the experiment. It could be that the participants wished to keep the bribe intentionally *under the table* as a strategy to keep the money. We claim this because the Error in Wins and the Error in the number of icons described is not statistically significant in our analysis. Hence, the participants were aware about what was happening during the experiment and still didn't mention anything about the unasked help via cheating. Apparently people knew that they could have this extra money and describe fewer icons but they still kept quiet. Those who reported in the interview the *strange behaviour* did not make any moral judgement towards the conduct of the robot. This can be linked with the work of Kahn et al., who suggest that people tend to keep robot secrets if the robot is in the room with the experimenter during the interview [75]. That could be explained in terms of the moral moral accountability given to the robot as is suggested by Kahn et al. [74]. . In addition to this, 36% (11 of 30) participants reported the bribe in the feedback of the questionnaire on the computer, but not in moral terms. Five of them reported feeling confused about the robot behaviour, two interpreted the bribing behaviour as kindness, two expressed disappointment and two had a desire to reciprocate the help.

We mentioned that the briber robot was also rated higher in the likeability score of the Godspeed scale. Apparently the unexpected behaviour of the robot increased the likeability

scores. This is in line with the results of Short et al., who reported that people feel more engaged with their cheater robot playing RPSG compared with the robot playing normally [142]. However, we must consider that the robot used in Short et al. study cheated at the expense of the participant whereas our robot cheated to allow the participant wins.

3.8 Conclusions

In summary, we designed a pioneer experiment in terms of experimental robot morality. Our results suggest that people are keen to reciprocate help to robots when they ask for a favour. However, they reciprocate less with the bribing robots compared with honest ones. Interestingly our bribing robot scored higher in likeability compared with the control condition. Apparently people felt attracted to its unexpected behaviour. In terms of the main effects, the use of direct speech or indirect speech was not significant for the participants. However there was a significant interaction effect between the play style and the speech used by the robot. Direct speech worked better in the no bribe condition and indirect speech in the bribe condition.

Additionally humans tended to maintain secrecy about the briber robot's behaviour in the interview but communicated more openly about the bribery in the online questionnaire. However, they did not report the bribe in moral terms; they were confused by the robot behaviour, interpreting its bribe as a kindness or malfunction. Only two of them expressed any obligation to reciprocate towards the robot. In other words, the robot was persuasive and subtle enough in bribing the people that some of them could have been unaware of it. Conversely participants could not report the bribe in order to keep the money.Seemly that people don't think in robots in high moral terms or want robots acting in moral terms as is suggested by Voiklis and Malle [153] and Johnson and Axinn [71].On the other hand Malle et al. [96], and Malle [97] work suggest possible solutions to the development of autonomous robot algorithms capable of moral decisions and moral limitations for robot autonomous behaviours [31].

Our work complements the existing body of related *HRI* research in reciprocity incorporating a quantitative approach through the RPSG to measure bribery as one of the dark sides of the reciprocity in the Human-Robot Interaction field. As a result, our study has an impact on the future design of human-robot interactions. We suggest that robot technology would not totally inhibit the natural human reciprocal behaviour in a bribery context. However the fact that humans will reciprocate less with a bribing robot than an honest one could have future consequences for the development of robot behaviours. Robot designers should consider that humans reciprocate toward robots in different contexts including a bribery scenario,

but significantly less than they would towards humans as Sandoval et. al., shows in [137]. Hence, it would be useful to conduct a future study with human bribers instead of robots playing RPSG to confirm our statements. Additionally, in the future humans should learn where are the moral boundaries for robots, and robot designers should forecast what kind of robot behaviour is appropriate according to the moral conventions. Furthermore, robot designers should improve the behavioural design of their artifacts so that human users can easily perceive when the robot is replicating a dishonest human behaviour and act according to the situation.

3.8.1 Limitations and future work

Due to the novelty of the current study, I propose the use of higher bribes offered by the robots in the decision games in future studies. Also, the encoding and analysis of the qualitative information by neutral reviewers is required to rank the responses of the participants objectively. According with Wallach [154], it is very possible the construction of moral robots. Hence, Robots with different embodiment and aesthetics could also be necessary to compare their influence in the human response.On the other hand, other experimental setups can be proposed as the suggested by Roizman et al. [132] and Ullman and Malle [152].

As a limitation of our work, we can mention that participants were curious about the capabilities of the robots because only 20% had previous experience with robots. Also, the use of just one type of robot is a considerable limitation since the embodiment, degree of anthropomorphism and voice could be factors affecting the users. Further statistical analysis should be performed with a bigger sample, normal distribution, homogeneity and not outliers. Also further studies should be performed considering related variables as trust[152] and secrecy.

Chapter 4

Likeability and Benefits of Robot Reciprocal Strategies

The most interesting characters keep us hooked. Not likeable ones! Iago, Shylock, Darth Vader - are they likeable? Do you want to invite them to dinner?
Alison Owen

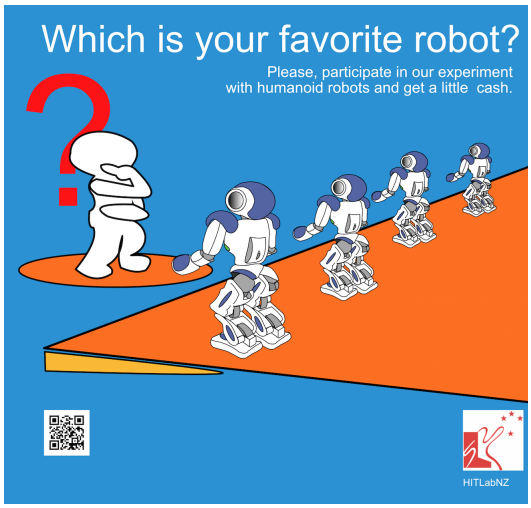


Fig. 4.1 Art developed for recruitment.

The material of this chapter has already been submitted as the paper: Sandoval, E. B., Brandstetter, J., Utku and Bartneck, C. (2016c). Measurement of reciprocal strategies in human robot interaction and their likeability using the alternated repeated ultimatum game. to the International Journal of Social Robotics, 2016, [138].

4.1 Summary

In this chapter we discuss how likeability in robots is an important multi-factorial phenomenon that has a strong influence on long-term relationships. One of the factors that affect likeability is the reciprocal response of the people towards the actions of the others. In *HRI* , likeability is constrained to behavioural variables since similar robots have identical physical embodiment. Our general research question in this study is: What type of robot reciprocal behaviour is better for humans when the robot's decisions effects them? We designed a between/within 2x2x2 experiment in which the participant plays Alternated Repeated Ultimatum Game (*ARUG*) for 20 rounds with NAO robots using four different reciprocal strategies. The factors were the offers (offer or inverted offer) and the acceptance of the offers (be reciprocal or inverted reciprocal). We had two between groups; humans starting the game and robots starting the game. We measured the money granted to the participant, the number of reciprocations of the game and the reciprocal offers during the game in order to compare with the likeability scores and preference over the robot strategies.

4.2 Introduction

Likeability is associated with friendly, cooperative and pro-social behaviours [28] such as extroversion, agreeableness, and lack of over-conscientiousness [45, 103]. Moreover, likeability is a very complex phenomenon involving behaviours, manners, perceived intelligence, similar socio-cultural context, interests, and even physical attractiveness, acceptability and popularity. For instance, a person is considered likeable when he or she is emotionally well-adjusted and he or she can be engaged into high-quality relationships.

The future acceptance and popularity of social robots will depend on their likeability. The measurement of likeability in robots is mostly associated with their degree of anthropomorphism [66] and the design of the embodiment. However, the likeability of state-of-art robots cannot be based on unique physical features. Robots of the same model will be mass produced; therefore, they will have identical physical embodiments. Sooner rather than later, they may lose their novelty effect and their appearance might become ordinary. Hence the likeability of the robots will be determined mostly by their behaviours towards humans.

We assume that people will find robots likeable depending on three main conditions which are independent of their external characteristics: A) To what extent is the robot useful to the user? In other words, how successfully can the robot perform the tasks that users expect? B) How does the robot behaviour match the interest and personality of the users? For instance, does the robot present slightly unexpected behaviour to keep the attention of a

curious human, and more predictable behaviour for users who prefer routine? C) How does the human-robot interaction benefit the human materially and emotionally? This third point leads us to ask if humans would drive the robot's behaviours based on their own self-interest or show reciprocal behaviour towards the robots if they received benefits from doing so.

Several studies have been performed on the acceptance and likeability of the robots [106]. However, most of theses studies focus on the natural human-robot interaction and do not consider the material benefits to the human and the reciprocally beneficial human-robot interaction which is an important factor for the likeability of the robots. Furthermore, some of the research measuring likeability in *HRI* has been performed using images, videos of robots or static robots instead of real interaction with a robot [10, 18, 107].

On the other hand, very recently, some studies reveal that humans tend to be reciprocal towards robots and computers playing Decision Games or when they ask for help [55, 57, 112, 137]. Furthermore, humans try to reciprocate to robots even when this breaks the social rules as our study in Chapter 3 describes. These facts can possibly be attributed to the likeability of the robots. The work of Sandoval et. al. surprisingly found that the users find the anti-social behaviour of a robot briber likeable [137]. Hence, we consider it necessary to investigate how different robot behaviours, particularly reciprocal behaviours, could have an effect in the robot-likeability.

We consider that reciprocity will determine how deep and meaningful are the interactions between the humans and the agents. Although we cannot claim that robots and humans will develop such deep relationships as friendship or love; we consider that in the future robots with a more engaging, interesting and likeable behaviour will have more chance to be accepted and be popular among the humans despite their lack of physical attractiveness[151, 157].

Hence, in this paper we aim to describe quantitatively the relationships between robot-likeability defined in the Godspeed questionnaire [9] and the reciprocal interactions between NAO robots and humans playing 20 rounds of Alternated Repeated Ultimatum Game (*ARUG*). We also measure the correlation between four different reciprocal strategies used by the robot playing ARUG and the likeability scores and preference ranking.

4.3 Literature Review

The use of the term "likeable" is broad. Defined shortly as: *easy to like and having pleasant or appealing qualities* [44] it allows several uses of the term. Extensive research has been done on likeability in Human-Human Interaction. Some of this can be analogous to robot-likeability research in that people try to find the way to be likeable when they are part of

a new group or in a new environment. For instance likeability in adolescents [45, 103]and social groups living in unfamiliar environments have been investigated[65].

Likeability is an important topic in *HHI* and *HRI* because humans tend to establish our relationships based on how we like (or dislike) certain kind of persons. When humans start a friendship or a romantic relationship they tend to do it based on likeability criteria. Over time, physical attractiveness and other shallow factors tend to become less important in the building of a relationship, and focus more on the emotional and material benefits mutually obtained.

However, likeability can be a contradictory phenomenon. Apparently people can find behaviours that are not necessarily reciprocal, cooperative and mutually beneficial attractive. The nicest behaviour of a person is not necessarily the most likeable for others; sometimes it is perceived as boring. Conversely, in certain cases, a subject can be aggressive, arrogant and manipulative, but despite that, people might still find them likeable [28]. Public figures such as rock-stars, athletes and politicians sometimes show rude or even disgusting behaviour but they still fascinate the public. For instance, Justin Bieber, Dennis Rodman, Kim Khardashian, Miley Cyrus, Paris Hilton, Kanye West and many, many other celebrities in different fields.

The question we raise about likeability in the domain of *HRI* is: Can these facts be translated to a Human-Robot Interaction? In other words: Should robots show unlikeable behaviour in order to be liked? Previous studies have shown that robots showing unexpected behaviours or even unacceptable behaviours received higher scores in likeability as I discuss in Chapter and [136]. Apparently our reciprocal relationships with robots are almost as complex as our relationships with other humans.

4.3.1 Likeability and reciprocity

In previous chapters I have mentioned that Fehr and Gachter discuss reciprocity in terms of positive and negative reciprocity *"...in response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self-interest model; conversely, in response to hostile actions they are frequently much more nasty and even brutal..."* [49, 50, 53]. Likeability and Reciprocity are strongly connected; if we consider somebody friendly and pleasant it is because generally we receive a reciprocal treatment from this person or agent. In order to measure reciprocal behaviours related with likeability in *HRI* , we use the insights of Game Theory.

The Repeated Ultimatum Game has offered a valuable instrument to measure different psychological and economic measures. For instance, Burnell et. al., have researched the optimal strategies without fairness when the Repeated Ultimatum Game is played [23] and Oosterbeek found common behavioural patterns regardless of cultural differences in a meta

analysis of Repeated Ultimatum Game [119]. Besides individual differences related with reputation [145], and attractiveness [143], the strategies displayed during the Ultimatum Game (UG) have been studied in depth in the economics field. These concepts are strongly linked with the concept of likeability that we use in this study. However, in our case likeability is more related to the robot behaviour and its reputation along the game rather than its physical appearance, anthropomorphism or embodiment.

Robot designers try to implement highly cooperative behaviours in robot but these are not necessarily the best solutions in terms of keeping the attention and interest of the user to interact with the robots socially. We consider that a reciprocal behaviour in robots can offer better results in terms of an effective, useful and engaging social interaction. However several studies have been done using decision games (Ultimatum Game, Prisoner's Dilemma and Rock, Paper Scissors Game as measurement instruments [91, 114, 137, 137, 142].

4.3.2 Alternated Repeated Ultimatum Game

Ultimatum Game is a well-known game used very often in Behavioural Economics experimental research [49]. In the original version, a Proposer decides how to distribute a certain amount of money and the Acceptor can decide whether to accept the distribution and both of them can keep the money. If the acceptor rejects the offer both of them lose the money. In our proposed configuration of Alternated Repeated Ultimatum Game ($ARUG$) the players alternate roles every round. For instance, player 1 is Proposer in round 1 and Acceptor in round 2 and so on. Also the robot and participant have 9 predetermined options to distribute the dollars between them . The options are: Human 10 dollars:Robot 90 dollars, Human 20 dollars:Robot 80 dollars,...,Human 90 dollars:Robot 10 dollars. An additional condition exist when the robot starts the game. In this case the robot initiates his game with an offer of 50 dollars Human:50 dollars Robot.

4.4 Research questions

The aim of the experiment is to analyse the participant's response in terms of robot's likeability (RL), participant's reciprocal decision (PRD), participant's reciprocal offer (PRO), and participant's profit (PP). We describe these measurements in section 4.5.6. Participants were expose to two factors: robot's reciprocal decisions (RRD) and robot's reciprocal offer (RRO) in $ARUG$ Game. Additionally , there is a between condition called Group G that describes if the participant or the robot starts the $ARUG$. In order to evaluate our aim we propose four research questions:

1. Is *RL* significantly affected by *RRD*, *RRO* and *G* individually or interactively?

2. Is *PRD* significantly affected by *RRD*, *RRO* and *G* individually or interactively?

3. Is *PRO* significantly affected by *RRD*, *RRO* and *G* individually or interactively?

4. Is *PP* significantly affected by *RRD*, *RRO* and *G* individually or interactively?

5. What is the correlation between *RL* and *PRD*, *PRO*, *PP*?

6. Do participants rank to robots significantly differently depending on the robot's factors?

4.5 Method

We conducted a mixed between/within 2x2x2 factors experiment in which the between factor is *G*, in other words, the starter of the session is human or robot and the within factors (2x2) in the 20 rounds of the $ARUG$ are *RRD* and *RRO*. *RRD* has two conditions: Tit for Tat's decision (TfT) and Inverse Tit for Tat's ($I-TfT$). Similarly, *RRO* has two conditions: Reciprocal Offer (*RO*) and Inverse Reciprocal Offer ($I-RO$).

TfT means that the robot follows the decision of the participant. For instance, if the participant accepted the robot's offer in round X, the robot will accept the participant's offer in round X+1. Conversely, $I-TfT$ consist in reject the offer in the current round if the participant accepted the offer in the previous round.

In *RRO* factor, *RO* condition means that the robot matches the participant's offer in terms of distribution. For instance, in *RO* if the participant offers a distribution such as: Human 10 dollars: Robot 90 dollars, the robot offers the same reciprocal distribution distribution in the next round; it means Human 90 dollars: Robot 10 dollars. In $I-RO$ the robot offers a non-reciprocal distribution. To illustrate, if the participant offers a distribution such as: Human 10 dollars: Robot 90 dollars, then the robot will offer an inverse distribution such as Human 10 dollars: Robot 90 dollars in the next round. See figure 4.2.

These factors are perceived by the participants as individual strategies of four robots (A, B, C, D) in the experimental conditions. We named the robots in this way in order to make them easier to remember for the participants. These strategies deployed by the four robots were the result of 4 combinations of the *RRD* and *RRO* conditions; TfT x *RO* is Robot A, $I-TfT$ x *RO* is Robot B. TfT x $I-RO$ is Robot C and $I-TfT$ x $I-RO$ is Robot D. See Table 4.1.

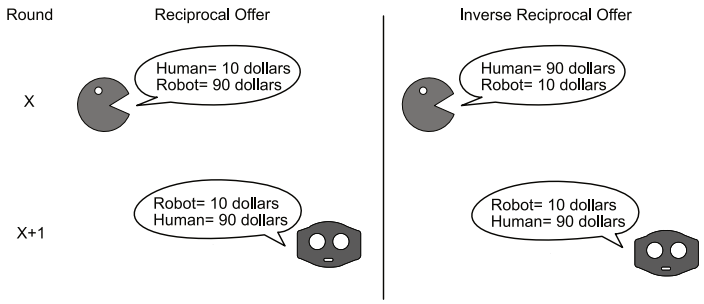


Fig. 4.2 The figure illustrates the differences between *RO* and $I-RO$ in two consecutive rounds. In $I-RO$ if the participant is selfish, the robot reciprocates generously.

Table 4.1 The Four factors used in the experiment.

Robot / Human Starting		**RRD**	
		TfT	*I-TfT*
RRO	*RO*	Robot A	Robot B
	I-RO	Robot C	Robot D

4.5.1 Experimental Setup

We rotated the order of the robots using Latin square method. Interaction between the participant and the robot involved both visual and audio communication. All four robots showed the minimal level of verbal interaction and animacity to minimise emotional impact on likeability for different robots. The robot made an offer to the participant in speech as well as pointing to a card that displayed its offer and received the human's response in speech, and the human players made their offer by showing a card displaying their offer to the robot. Apart from relaying their offer/response and guidance, robots also verbally rephrased participants' actions. For instance, after a participant offered Human 70 dollars : Robot 30 dollars, the robot would say "You offer me 30. Ok, I accept it".

4.5.2 Materials

We used one NAO robot, presented under the disguise of four different robots to participants. Experiment layout had an "Accepted" area and a "Rejected" area for the offers to be put into accordingly. A fixed layout of cards with offer rates was placed before the robot, to which it pointed with its finger to indicate its offer. Twenty units of cards for each offer rate were placed in a similar fashion in front of the participants, and were used for making offers to the

Fig. 4.3 Setup of the Experiment. The participant can choose from nine different options and the robot can point out the options.

robots and also for tracking the accepted/rejected amounts. A laptop was placed on a nearby desk for the online questionnaire. See Figure 4.3.

4.5.3 Process in Human Starting Condition

In both conditions, after introducing the mechanics of the experiment to participants, we started the experiment and discretely observed the first 2 rounds from outside of the room to make sure the participant was not having technical problems, and then we left the room. After each session we came back to change the robot and calculate the results of the sessions to compare them with data recorded in the robot and left again so that participants would not feel pressured by having an observer. Participants filled an online survey with the Godspeed questionnaire after each experiment and a comment section regarding their opinions of the robot. After all four sessions participants filled out the ranking about how much they liked each robot. At the end of the experiment participants were compensated by 0.03% of their accumulated symbolic earnings which ranged between $6.00 and $13.00. This experiment was approved by the Human Ethics Committee of the University of Canterbury [HEC APPLICATION 2015/36/LR-PS].

We performed individual sessions of four *ARUG* games. Participants were welcomed and led into the experiment room to receive a brief description of the experiment. After reviewing

and signing a consent form, they were asked to fill out an online questionnaire that gathered demographic data including their previous experience with robots and we provided an ID number to each participant. Then participants were shown a short film clip that demonstrated the experiment process. When they were ready, the robot was activated through its feet bumpers, after which it asked the number of the participant, who replied verbally. After each session, the experimenter came to count the cards, re-stacked them and pretended to replace the robot with the next one. At the very end participants were told their total score and thanked for their participation. Then any questions were briefly answered. See Figure 4.4 to see the simplified flow of the rounds.

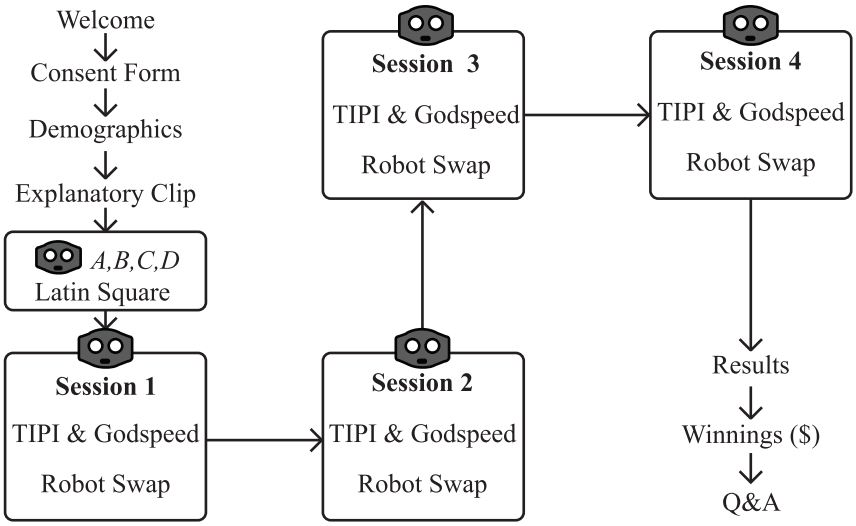


Fig. 4.4 Experimental proceduree

A NAO robot was introduced as the participant's opponent in *ARUG*. The robot wore a tag that displayed "A","B","C" or "D" to emulate the perception that the participant was facing four different robots (whereas we used a single robot and reprogrammed it between sessions). The robot asked for the ID number to start the session in each condition. Once the session started the robot requested the participant to take the first turn, and asked the participant to show the card that displayed the offer they wanted to give. By default all four robots were programmed to accept the first offer to prevent participants from identifying its action pattern on the first round. Starting from the 2nd round the robot started its programmed reciprocation patterns. We designed in this way in order to be consistent with the assumption of the cooperative behaviour of social robots. After each session the robot was taken out of the room, and while the participant filled out the survey, the robot was reprogrammed for the next reciprocation pattern and its tag replaced accordingly, then represented to the participant as their new opponent.

Each round the robot announced the number of the current round, and then if it was the robot's turn to offer it pointed to the proper card and asked whether the participant accepted or not, using the speech recognition system recognising a Yes or No. If it was participant's turn to offer, the robot told the participant to hold the card bearing the offer, and then it gave its response based on its reciprocation pattern using its vision system. At the end of the session of 20 rounds, the robot announced that the session was over. The participant then was asked to fill out a survey on their opinion on the robot and their perceived earnings. After completing the final survey they were given the amount of their earning on that session. See Fig. 4.5 to see the experimental process per participant.

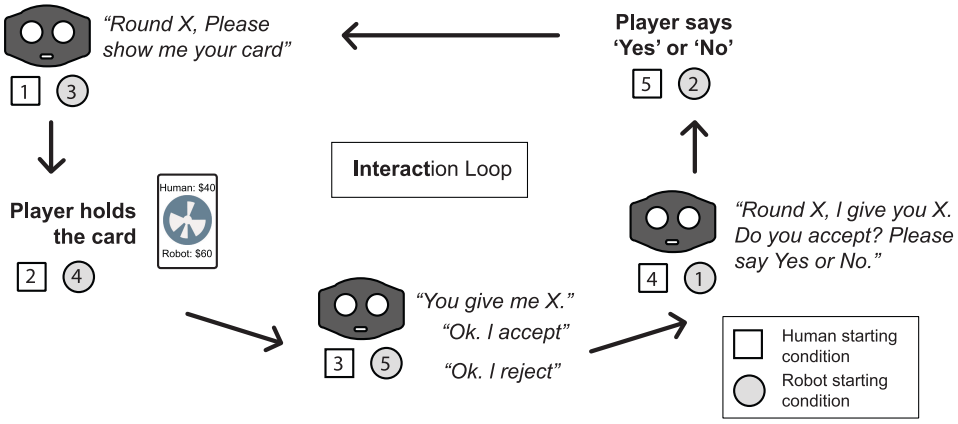


Fig. 4.5 Initialising the game depending if the human or the robot starts.

4.5.4 Process in Robot Starting Condition

In robot starting condition, once the experiment started the robot informed the participant that it would take the first turn, and offered Human 50 dollars: Robot 50 dollars as default in all four sessions to prevent the participants from identifying its action pattern in the first round. Starting from the 2nd round the robot started the programmed reciprocation patterns.

4.5.5 Participants

We contacted participants via university boards, dedicated websites to recruit participants, and Facebook groups in the city. After disposing of the data of sessions which were not carried out successfully due to human error or a robotic malfunction, we had 38 participants in our experiment: 20 in the robot starting condition and 18 in the human starting conditions. Half of the participants were male. 42% of the participants had prior experience in interacting with robots in demonstrations and studies. 5% were high school graduates, 42% were currently in

college, 21% had college/university degrees, 13% were currently in graduate or professional school, and 18% had graduate or professional degrees. 68% of the participants currently had jobs. 37% were from Oceania (Australia and New Zealand and other countries), 29% from Asia (China, India, Japan and others), 18% were from Europe and 16% from the North and South America. The average age was 25 years old (SD=6.99).

4.5.6 Measurements

The measurements in the experiment are: *RL*, which is an item of the Godspeed questionnaire series [9], participant's reciprocal decision (*PRD*), means that the participant follows the behaviour of the robot in the immediate next round, participant's reciprocal offer (*PRO*), means that the participant matches the offer of the robot in the immediate next round, and participant's profit (*PP*) obtained by the participant in each condition.

4.6 Results

We performed a three-way mixed ANOVA (2x2x2) in which the between factor is *G* and the within factors are *RRD* and *RRO*. The measurements were *RL*, *PRD*, *PRO*, and *PP*. See measurements, interaction effects, main effects, means and standard deviations of each measurement in Table 4.2, 4.3, 4.4 and 4.5.

The first research question investigated the effect of *G*, *RRD*, *RRO* on *RL*. There is a statistically significant three-way interaction effect between *RRD*, *RRO*, and *G*, (F(1, 36) = 6.072, p=0.019). The outliers were kept in the analysis because they did not materially affect the results as assessed by a comparison of the results with and without the outliers. There was a significant two-way interaction (F(1,19)= 4.452, p=0.048) between in *RRD* and *RRO* appeared in the human condition but not in the robot condition (F(1,17)=1.930, p=0.183). There is a significant simple main effect (F(1,19)=4.902, p=0.039) of *RRD* in the human group condition and a significant main effect of *RRD* in the robot group, (F(1,17)= 10.742, p=0.004). See Table 4.2 for interaction effects, means and standard deviations.

The second research question investigates the effect of *RRD*, *RRO* and *G*, on *PRD*. A statistically significant three-way interaction between *RRD*, *RRO* and *G* that affects *PRD* was found, (F(1, 36) = 12.665, p=0.001). There is a significant two-way interaction between *RRD* and *RRO* in the human condition, (F(1,19)=15.092, p=0.001). However, there is a non significant two-way interaction between *RRD* and *RRO* (F(1,19)=1.294, p=0.271) in the robot condition. There is a significant simple main effect of *RRD* in the human condition, (F(1,19)=5.608, p=0.029). There is a significant simple main effect of *RRO* in the human

Table 4.2 Interaction effects, main effects, means and standard deviations of robot's likeability

Measurement	Three-way interaction	F(1, 36) =	Human*TfT*RO	Human*TfT*I-RO	Human*I-TfT*RO	Human*I-TfT*I-RO
RL	G*RRD*RRO	6.072, p=0.019	3.66(0.85)	4.12 (0.81)	3.61(0.85)	3.51 (1.08)
			Robot*TfT*RO	Robo*TfT*I-RO	Robot*I-TfT*RO	Robot*I-TfT*I-RO
			4.01(0.69)	3.97 (0.79)	3.16 (0.85)	3.49 (0.84)
RL Human	Two-way interaction	F(1,19)=				
	RRD*RRO	4.452, p=0.48				
	Main effects	F(1,19)=				
RL Human	RRD	4.902, p=0.039				
		F(1,17)=				
RL Robot	RRD	10.742, p=0.004				

Table 4.3 Interaction effects, main effects, means and standard deviations of participant's reciprocal decision (*PRD*).

Measurement	Three-way interaction	F(1, 36) =	Human*TfT*RO	Human*TfT*I-RO	Human*I-TfT*RO	Human*I-TfT*I-RO
PRD		12.665, p=0.001	7.25(2.17)	8.2(2.04)	5.35(1.73)	3.0(1.65)
			Robot*TfT*RO	Robo*TfT*I-RO	Robot*I-TfT*RO	Robot*I-TfT*I-RO
			9(0)	6.94(2.6)	2.89(2.14)	1.78(1.9)
	Two-way Interaction effects	F(1, 19) =	TfT*RO	TfT*I-RO	I-TfT*RO	I-TfT*I-RO
PRD Human	RRD*RRO	15.092, p=0.001	7.25(2.17)	8.20(2.04)	5.35(1.73)	3.0(1.65)
	Main effects	F(1, 19) =	RO	I-RO		
	RRO	32.589, p<0.001	6.1(2.825)	5(3.36)		
			TfT	I-TfT		
	RRD	5.608, p=0.029	7.8(2.1)	3.3(2.245)		

condition, (F(1,19)=32.589, p<0.001). Besides, there is a significant simple main effect of *RRD* in the robot condition, (F(1,17)=11.018, p=0.004) and there is a significant simple main effect of *RRO* in the robot condition, (F(1,17)=104.171, p<0.001). See Table 4.3 for interaction effects, main effects, means and standard deviations. Outliers were not removed from the data.

In the third research question we investigate if *RRD*, *RRO* and *G* affects *PRO*. We found that there is not an significant three-way interaction effect (F(1,36)=0.824, p=0.370). There is a statistically significant main effect of *RRO*, (F(1, 36) =4.151, p= 0.049). There is a statistically significant main effect of the *RRD*, (F(1, 36) =8.775, p= 0.005). Besides, there is a between subject main significant effect of *G*, (F(1, 36) =8.137 p= 0.007). See Table 4.4 for the means and standard deviations.

Table 4.4 Interaction effects, main effects, means and standard deviations of participant's reciprocal offer (*PRO*).

PRO	Main effects	F(1, 36) =	RO	I-RO
	RRO	4.151, p=0.049	2.7(2.49)	1.6(1.96)
			TfT	I-TfT
	RRD	8.775, p=0 .005	2.4(2.733)	1.9(1.763)
			Human	Robot
	G	8.137 p= 0.007	1.575(1.833)	2.79(2.6)

Table 4.5 Interaction effects, main effects, means and standard deviations of participant's profit (*PP*).

Measurement	Two-way Interaction effects	F(1, 36) =	TfT*RO	TfT*I-RO	I-TfT*RO	I-TfT*I-RO
PP	RRD*RRO	34.006, p<0.001	752.63(251.93)	1219.21(387.05)	567.63(97.24)	707.89(122.48)
	Main effects	F(1, 36) =	RO	I-RO		
	RRO	66.515 p<0.001	660.1(211.297)	963.6(384.106)		
			TfT	I-TfT		
	RRD	76.536 p<0.001	985.9(400.46)	637.8(130.57)		

Table 4.6 There are significant moderate and weak correlations among *RL*, *PRD*, *PRO*, and *PP*.

p<0.02	PRD	PRO	PP
RL	0.308	-0.225	0.226
PRD			0.513

In terms of *PP*, there is not a statistically significant three-way interaction between strategy, offer and group, (F(1, 36) = 0.053, p=0.819). Outliers were not removed from the data. However, there is a statistically significant two-way interaction between *RRD* and *RRO*, (F(1, 36)=34.006, p<0.001). A statistically significant main effects (F(1, 36) =76.536 p< 0.001) of *RRD* were found. Besides *RRO* present a significant main effect (F(1, 36) =66.515 p< 0.001). See Table 4.5 for the means and standard deviations.

In order to answer our fifth research question, we determine the Spearman's correlation between the *RRD*, *RRO*, *PP* and *RL*. Preliminary analysis showed the relationships to be monotonic, as assessed by visual inspection of the scatter-plots. There was a positive moderate correlation between *PP* and *PRD*, rho (152) = 0.513, p<0.0001 and a positive moderate correlation between *RL* and *PRD*,rho (152) = 0.308, p<0.0001. Besides, there was a positive weak correlation between *RL* and *PP*, rs(152) = 0.226, p < 0.005 and a negative weak correlation between *RL* and *PRO*, rs(152) = -0.225, p<0.005.See Table 4.6.

Finally, for the sixth question, in order to determine the favourite robots, we asked the participants to rank them. We conducted a Chi square goodness-of-fit test to determine whether participants ranked one of the four robots significantly higher than the other robots. The minimum expected frequencies were 9.5 for the general ranking, 5 for the Human starter group and 4.5 for the Robot starter group. The robot condition TfT x *RO* (pure reciprocal) was ranked highest followed by TfT x I_RO condition. However, all Chi square values for the four different robots are not not significant likely due to the size of the sample. See table 4.7.

Table 4.7 Ranking of robot reciprocal conditions

TfT x RO and $I-TfT$ x RO received the best rankings due to probably the consistent reciprocal strategy and the economic reward received by the participant respectively.

General Ranking of the robots				
Favourite	**1st**	**2nd**	**3rd**	**4th**
TfT x RO	13	13	7	5
TfT x $I-RO$	9	9	9	11
$I-TfT$ x RO	8	9	11	10
$I-TfT$ x $I-RO$	8	7	11	12
Human Starter of ARUG				
TfT x RO	5	8	3	4
TfT x $I-RO$	5	5	6	4
$I-TfT$ x RO	5	5	5	5
$I-TfT$ x $I-RO$	5	2	6	7
Robot Starter of ARUG				
TfT x RO	8	5	4	1
TfT x $I-RO$	4	4	3	7
$I-TfT$ x RO	3	5	5	5
$I-TfT$ x $I-RO$	3	4	6	5

4.7 Discussion and Conclusions

The robot's likeability is affected by the three-way interaction effect of *G*, *RRD*, and *RRO*. Hence, a two-way ANOVA was performed by separated groups. A two-way interaction effect between *RRO* and *RRD* was found. The robots displaying a reciprocal decision TfT were rated higher in likeability than the robots using a inverse reciprocal offer $I-TFT$. Indeed, the robot in the TfT and $I-RO$ condition had a higher likeability score than the other robots in the human condition. In this case $I-RO$ is beneficial for the robot in the TfT condition but not for the robot in TfT and RO condition (pure reciprocal). In the case of the robot condition, there there is a main effect of *RRD* such that TFT (M=3.98, SD=0.73)lead to higher scores of likeability than $I-TfT$ (M=3.32, SD=0.85).In the robot group the robot in the TfT and $I-RO$ condition also has higher scores. The likeability of the robot due to the TfT and $I-RO$ could be explained by the unexpected behaviour of the robots towards the participants and the nature of $I-TfT$ that reciprocate low offers with higher offers as is explained in the next paragraph. The study performed in chapter 3 shows similar results in terms of the likeability of robots performing unexpected behaviours even when these behaviours were breaking the social rules. Moreover, these results slightly match with the results of the ranking of the robots after all the experimental sessions. The favourite robots in

the ranking were firstly the robot in TfT and *RO* (pure reciprocal) condition and then TfT and $I-RO$ condition . Probably robot in TfT and $I-RO$ was perceived as a generous robot and that is why people like it. They didn't expect that offering low pays then they received higher pays from the robot during the $ARUG$. Participants liked the unexpected economical benefit and "nice" behaviour of the robot. On the other hand the pure reciprocal could be perceived as a easier to understand.

In terms of participant's reciprocal decision, *PRD* a significant three-way interaction effect existed. Then, a two-way ANOVA was performed for each group. A two-way interaction effect between *RRO* and *RRD* was found. Participants reciprocate more towards the robots in the TfT condition than in the $I-TfT$ in the human group. See Table 4.3 for means and standard deviations. Participants reciprocated more towards the robot in the TfT and $I-RO$ condition in the same group. This results are in line with our results in chapters 2 and 3 in terms that the Norm of Reciprocity [63] applies in HRI. People tend to reciprocate towards robots that show an evident reciprocal behaviours. Moreover, they naturally reciprocate more towards the robot in TfT and $I-RO$ because it offers higher economical benefits. The robot made higher offers when the participant offered little money. None cases of humans offering high amounts of money to receive little money from the robot appeared during the experiment. In the case of the the robot group, *RRD* had a main effect in the decisions of the participants. They reciprocate more to the robot in TfT condition, (M= 7.97, SD=2.09) than the robot in $I-TfT$ (M=2.3, SD=2.07). Similarly, *RRO* had an impact in *PRD* in the robot group. Participants reciprocate more frequently when the robot used a reciprocal offer in *RO* condition (M=5.94, SD=3.44), than when the robot was using $I-RO$ (M=4.36, SD=3.44). Similarly to the human group, reciprocal strategies play a role that lead to think that the Norm of Reciprocity rules the reciprocal behaviours in HRI. Moreover, the use of simultaneous reciprocal different strategies has a very defined outcome in terms of *PRD* and *RL*.

In terms of *PRO* there are not interaction effects at all. There is a main effect of *RRO*. Participants reciprocate the offer of the robot more frequently in *RO* (M=2.7, SD=2.49) than in $I-RO$ (M=1.6, SD=1.96). There is also a main effect of *RRD*; participants reciprocate the robot's offer more often in the TfT condition (M=2.4, SD=2.733) than in the $I-TfT$, (M=1.9, p=1.763). Besides, the group, makes a significant difference in *PRO*. Participants reciprocate the offer less (M=1.575, SD=1.833) when they start the $ARUG$ than when is the robot who starts the game, (M=2.79, SD=2.6). Apparently, the robot is capable to establish a reciprocal pattern when it starts the game that is easy to follow by the participant.

In terms of *PP*, there was not three-way interaction effect. However there is a two-way interaction effect between *RRO* and *RRD* that can be explain with the main effects.

In *RRO* condition participants had a higher profit whit the robot in the $I-RO$ condition (M=963.6, SD=384.106) than in the *RO* condition, (M=660.1, SD=211.297). Similarly in *RRD*, participants had a higher profit with the robot in TfT condition , (M=985.9, SD=400.46) than with the robot in $I-TfT$, (M=637.8, SD=130.57). In other words, the combination of TfT and $I-RO$ are the most profitable for the participant. The combination of the reciprocal movements and negative reciprocal offers made the participant quickly notice that they can obtain higher profit if they keep making negative reciprocal offers (low offers) because the robot will offer high offers in the next round. The main effect of the *RRO* made more profitable the strategies that imply more *RRD*. For instance, a higher reciprocal offer coming from the robot makes it easier for the participant to accept it and do a reciprocal movement in the next round.

In terms of the correlations between *RL*, *PRD*, *PRO*, and *PP*, further studies are required due to the moderate and weak nature of the correlations.

Finally, participants ranked the robots at the end of the experiment. They had a general view of all the possible behaviours of the robots and freely decided their favourite robot in their own terms as we can note in their final comments. Although the chi square analysis does not offer significant results due to the size of the sample, the ranking gives some clue for future studies. People ranked TfT x *RO* as their favourite and TfT x $I-RO$ as their second favourite. In the case of TfT x *RO*, the Pure Reciprocal robot, this could be explained due to the fact that they could detect a reciprocal pattern easily compared to the other robots which had more unexpected behaviour. For TfT x $I-RO$ we observed a reciprocal pattern perceived as generous due to the higher reciprocal offer of the TfT x $I-RO$ when the participant made a low offer. This reciprocal strategy of TfT x $I-RO$ gave to the participants who noticed it early more money compared to the other strategies.

4.7.1 Conclusions

This study demonstrates that Humans accomplish the Norm of Reciprocity proposed by Gouldner [63] in the domain of *HRI* in terms of robot's likeability, participant's reciprocal decision, participant's reciprocal offers, and participant's profits. People like more the reciprocal robots such as TFT x *RO* and TfT and $I-RO$ conditions and obtained more benefits of the combination of strategies of *RRD* and *RRO*. TfT and $I-RO$ robot was likeable due to the unexpected behaviour bringing economical benefits to the participant.

This study is in line with the results of chapter 2 and 3. The Norm of Reciprocity rules the interaction of decision games in HRI in terms of *PRD* and *PRO*. When the human starts the interaction, participants reciprocate towards the robot that shows an evident reciprocal behaviours, specifically with the robot in the TfT and $I-RO$ condition due to the higher

economical benefits, when the participant offers little money to the robot. The robot made higher offers when the participant offered little money.

When the robot starts the interaction, participants reciprocate the offer (*PRO*) and the decision, (*PRD*) in the TfT and *RO* conditions, more often than when the human starts the interaction due to the robot establish a pattern easy to follow. Besides this robot starts the interaction with a 50%:50% offer that could be perceived as a fair offer. This perception could be the cause of the significant higher reciprocation towards this robot. This findings could be potentially useful in the future in order to design complex reciprocal behaviours for different social applications such as health-care, education or entertainment. Different layers of reciprocal behaviours could work together in order to keep the attention of the user and provide benefits by different means.

The participant's profit *PP*, is affected simultaneously by *RRD*, and *RRO* as main effects. Consequently participants obtain a higher profit with the robot in the TfT and $I-RO$ condition.

Although the people received a higher profit from the robot in TfT and $I-RO$ condition they ranked higher the pure reciprocal robot (TfT and *RO*) when they compare among all the robots. This is likely because this robot offer an easy understanding and predictable outputs during the *ARUG*. However in the experimental session participants find more likeable TfT and $I-RO$ robot condition. In other words, a likeable robot could be not necessarily the favourite robot when it is compared with other robots. However, robots showing in some extent a reciprocal robot; such as the TfT x *RO* and TfT x $I-RO$ would be more beneficial for the users than the robots that do not show a reciprocal behaviour.

4.7.2 Limitations and Future Work

Further studies are required in order to determine stronger correlations between likeability and reciprocity. In future studies a higher number of participants is required. In addition, the measurement and analysis of other items in Godspeed scale could be added to the study.

Chapter 5

Conclusions and Contributions

If you look at the field of robotics today, you can say robots have been in the deepest oceans, they've been to Mars, you know? They've been all these places, but they're just now starting to come into your living room. Your living room is the final frontier for robots.

Cynthia Breazeal

Fig. 5.1 Art developed for a conference.

In the previous chapters I have shown the fundamentals of how reciprocity works in *HRI* through a quantitative approach coming from Game Theory. My goal has been to offer to the reader extensive descriptions of my studies over reciprocity aiming at better understanding of the reciprocal phenomena. Additionally, I expect that these descriptions allow the repeatability of the experiments for people interested in the study reciprocity.

I would like to highlight that up until now, just few researches have been focused totally in reciprocity in the domain of *HRI* . Some studies have been done in the domain of *HCI* and virtual agents as I mentioned in section 1.3 but mainly focus on cooperation rather than reciprocity. Although there are similarities, *HRI* presents different challenges compared to *HCI* due to the intrinsic differences between computers and robots.

The novelty of my research mainly lies in the quantitative approach to a phenomenon that was suspected in *HRI* but not confirmed. Moreover, we quantify the reciprocity in *HRI* and now we know to what extent people reciprocate towards robots compared with how they do with humans, how reciprocity works in benefit of the robots and what are the most likeable reciprocal scenarios for humans.

The three main conclusions can be drawn as a result of the studies detailed in this thesis are:

1. Seemingly, the Norm of Reciprocity [63] under the operationalisation of the definition of Fehr and Gaechter [53] applies in HRI as studies in Chapter 2 and Chapter 4 suggest.

2. Cooperative, recognisable , but some unexpected robot behaviour is likeable to a certain degree for the users and beneficial for the robot as Chapter 3 and Chapter 4 suggest. However, these behaviours are not the most benefical HRI strategy.

3. More studies in reciprocity are required in more complex social scenarios as the *Limitation* section of each chapter suggest. The novel effect of the robot possibly affects the results of the experiments.

In the next sections, I explain in detail how I draw these conclusions. Besides,I consider that the findings of this thesis contribute to the evolution of robot technologies and will allow, in the near future, better designs for social reciprocal behaviours.To finish this thesis, in section 5.4, I dare to speculate on the impact of our findings in reciprocity for three main applications of reciprocity in *HRI* in the near future.

5.1 To what extent the Norm of Reciprocity applies in HRI?

Trying to respond the my first research question, to what extent do humans reciprocate towards robots? In Chapter 2 and Chapter 4, I state that people accomplish the Norm of Reciprocity when they play Prisoner's Dilemma with a robotic agent. However, they accomplish it to a lesser extent than in *HHI*. I can suggest that mainly they accomplish this Norm of Reciprocity with both humans and robots due to the reciprocal strategy displayed by the agents (Tit for Tat). This kind of reciprocal strategy could be used in different scenarios that allow an alternated interaction and the decisions of the participants to influence the next decision of his/her opponent. Hence, this interactions between robotic agents and humans could be mutually beneficial across long periods of time. However, we should give some considerations to this Norm of Reciprocity being applied to *HRI*. In scenarios like Ultimatum Game played once and influenced by the previous interaction in Prisoner's Dilemma the human response is different. Under these circumstances humans tend to exhibit a less fair behaviour towards the robotic agent compared to the human agent.

Considering these results, I can say that the setup of decision games can be translated to other more complex scenarios. I suggest a design rule of thumb for most of the common social interactions in the future: robots should exhibit a cooperative, fair and a little bit unexpected behaviour from the beginning of the interactions in order to trigger similar behaviour in humans.

5.2 Likable robot behaviours that could beneficial to the robot

The second research question deals with the fact; to what extent robots can use reciprocation to their own benefit and how likeable and unexpected is that? As we know now, the Norm of Reciprocity is accomplished in *HRI* to a lesser extent than in *HHI*. However, is there a way we could increase such reciprocation towards the robots? Furthermore, to what extent can robots use reciprocation to their own benefit?

The answer to these questions have implications in the future development of Artificial Intelligence/ Robotics and Virtual Agents Training Systems as I discuss in subsection 5.4.3. In Chapter 3, I describe a scenario where the robots are capable of triggering reciprocation from the people when they break the social rules and give extra rewards to the participants. This scenario is very possible in *HHI* and hence plausible for future robot behavioural designs. In this study, I conclude that people reciprocate also towards bribing robots again

accomplish the Norm of Reciprocity. However, they reciprocate less towards these robots who break the social rules of the games compared to honest robots.

Interestingly, participants found likeable the unexpected behaviour of the robot when the reciprocal patterns were not to distant of the original "tit for tat" strategy. Despite the fact the robot was bribing the participant in chapter 3 or rewarding the participant with illogical offers, the robot strategy was found likeable in a certain degree.

In other words, bribing robots are not the best method to increase the reciprocation towards robots. Hence, I suggest a second rule of thumb for robot design: we should keep robots honest. Although the likeability of these behaviours, we must find other ways to increase the human reciprocal response towards them. For instance, indirect language is a factor that increases the reciprocity towards briber robots compared to honest robots due to the similarities to *HHI*. However, the use of indirect language in robots is ethically questionable and impractical due to the secondary intentions between lines. In our study, just 10% reported the robot bribe in moral terms in the online survey. Participants don't judge them in moral terms as they judge a human following similar behaviour. This set of facts makes me wonder about several considerations of the ethical use of reciprocal robots.

5.2.1 Ethical Considerations of Reciprocal Interactions in HRI

Twelve years ago Fogg [58] discuss ethical considerations about persuasive technologies in his book *Persuasive technologies*. However he did not consider robots among these considerations. Now, we can consider robots as an emerging technology with powerful persuasive/reciprocal interfaces that will be commercially available in the next years. Hence, some comments should be made about ethical considerations in the use of reciprocal social robots.

Recent research in Neuroscience and Psychology shows that persuasive techniques used to increase the engagement in technologies such as mobiles games or online websites create undesirable human addiction towards these [37]. This engagement is fully based in negative reciprocal benefits between the user and the system. The user receives a pleasant reward of dopamine in exchange for spend money in these games. This phenomenon is known as *compulsion loop* which is used by certain paid video-games or gambling websites [25, 120] that could even create addiction to mobile phones, tablets and computers [161]. The user obtain an immediate emotional reward and the creator of the game an economical benefit. Besides, this reciprocal interaction between a digital device and an user is usually socially accepted. However these kind of methods creates moral dilemmas about their use in robots and the development of addiction towards them. For instance, How we can avoid addiction due to reciprocation to sex robots in mentally healthy human beings? How we can avoid

anti-social behaviour towards other humans due to a reciprocal robotic preference? Certainly, more research is required in these topics as is suggested in [96, 139].

I think that the implementation of mechanisms like the *compulsion loop* can be possible in robots. Certainly the patterns used in certain video-games in order to make them engaging could be done in robots. Specifically several strategy video-games use overlaped layers of decision games in order to keep the attention of the users in long periods and more important, get their money continuously in order to reach higher levels in the game. These well designed strategies plus the features of the robots as an attractive embodiment, and their animacity could make the robots not just engaging. Robots using inappropriate reciprocal strategies in order to persuade the users to use them can even create strong attachments as now happens with mobile phones. Konok et al. [83] claim that attachment to smart-phones can even generate anxiety. It is easy to see that a similar situation could be created in *HRI* . This attachment in the terms described by [17] in *HHI* could be similarly implemented in *HRI* and have significant negative consequences in *HRI* due to the lack of moral judgement over the reciprocal strategies of the robot. Turkle [151] describe some undesirable situations of attachment expressed by users about the future use of robot companions.

Therefore, I propose that further studies are required in order to determine the reciprocal strategies that generate the appropriated level of reward, attention and attachment in *HRI*. Moreover, these studies could help to minimise the moral dilemmas created in complex social *HRI* . However, I think that firstly it is necessary to determine the likeability of the reciprocal strategies of the robots. This likeability could be the variable that determines the future thresholds of reward, attention and attachment. For instance, if people find likeable certain reciprocal strategies due to the reward, then in a future interaction the user will keep the attention on the robot and then create an adequate level of attachment with it.

5.3 What are the Most Beneficial and likeable Reciprocal Robot Strategies?

In chapter 4, I describe a complex interaction between participants and robots through the Alternated Repeated Ultimatum Game (ARUG). This study tried to model more complex social *HRI* based in material benefits. Four different reciprocal strategies were used in the ARUG. These are related to higher benefits, and better perception of the robots.

I found that participants tend to reciprocate towards the most reciprocal robot when the robot starts the interaction. However but towards the most altruistic robots when the human starts the interaction due to the obvious pattern of higher benefits for the participant. Two

apparent contradictory phenomena happen in terms of likeability of the robot's reciprocal strategies. On one hand, there is a significant direct correlation between likeability and the pure negative reciprocal robot strategy (selfish robot). Two other strategies had a negative significant correlation. Similarly to our study of bribing robots, the likeability of the robots correlates to their unexpected behaviour. On the other hand, when the participants express their preference for the four different robot strategies, they state that the pure reciprocal robot and the altruistic robot represent their first and second favourite respectively. This preference could be attributed to the recognisable patterns that these robots show.

Considering these facts, I would suggest another rule of thumb of robot behaviour design: The behaviour of the robots should follow an easily recognisable reciprocal pattern and an occasional unexpected behaviour in order to increase its likeability.

5.4 Three Future studies of Reciprocity in HRI

More studies of reciprocity in complex social scenarios are required in *HRI*. Since I started my research in *HRI* some years ago, I have witnessed the dawn of different social robotic technologies aiming to help, assist and solve problems for the users. However most of these are pioneering work and not consolidated products. Although it is difficult to forecast the potential of these technologies, I dare to claim that the design of social robotic behaviours considering reciprocity as a main variable in complex social scenarios during the interaction are the key to achieve mature robot products. In this section, I do imagination exercises based in my personal research experience in order to propose future applications for reciprocity in HRI. However, it must be said that the proposed scenarios in this section have several limitations due to the practical intrinsic constraints of experiments based in decision games. Further longitudinal studies are required to describe long term *HRI*.

We are at a key moment in the history of robotics. We have projects as Buddy [14], Jibo [70], Pepper [144] Amazon Echo [2] that aimed to be consumer electronics products in 2015. With the exception of Pepper, they are glorified tablets or speakers, innovative as interfaces but not real robots under my definition of robot. Most of them have limited physical interaction with the users or environment or are totally static. In my opinion, a robot should be a machine that can perform some task in the physical world for the benefit of the user. The afore mentioned agents are clearly imperfect as social devices and require a lot of work to make them really useful for the users. On the other hand we have consolidated system such as "Siri" and "Ok Google" that will compete directly with other social technologies as wearable technology and the Internet of things to attract the attention of the costumers due to their impressive capacity to engage with the users, offer information and sort some of the

common problems of their daily life. Hence, robots should compete and perform better than these other technologies in order to become popular, useful and profitable.

As I said before, a insufficiently explored advantage of the social robots as a social technology is that they have the potential to do a physical work for the user as currently service robots like Roomba do now. Other technologies are not capable to do that and more research is required to found tasks that small robots can perform in houses, schools, hospitals and offices. Hence, I think if robot designers can conciliate the capability of the service robots to perform task for the users with well-developed social interfaces; then we will have a significant rise of commercial social robotics.

I speculate that the development of social robots will follow similar path than the personal computer industry in the 1980s. Firstly we will have researchers and early adopters using very basic social robot applications; and once we match the theory and practice of social robotics we will be able to add more sophisticated applications of social robotics. My experience as a robot behavioural designer and my observations of very diverse robot interactions through my demonstrations and experiments make me think that there are several immediate applications of the robot reciprocal behaviours.

5.4.1 Healthcare

Several studies have shown that robots have significant effect in the health of certain kind of patients [130, 131] even better than other technologies [98]. However the initial engagement between the patient and the robot tend to decline over time[21]. Possibly if a robot behavioural design considering reciprocal strategies is used, then we could obtain a higher engagement and alive interest in long' term *HRI* in healthcare.

As I mentioned in section 5.2.1, the design of complex behaviours based in decision games could help a lot to improve the *HRI* . I think that a *gamification* of the *HRI* as it is described by Deterding et al. [42] could be particularly productive if reciprocal robot strategies used in Health-care. Of course not all therapies are able to be a game but certainly the use of simple interaction with rewards can sort other problems in Healthcare. For instance, automated self-medication [101, 160] could improve if the gamification of reciprocal strategies were used to improve the interaction with the patient.

Keep in mind that complex social *HRI* in healthcare is limited to certain scenarios, personalities, diseases and disabilities present in the potential users. From my point of view, companion robots as Paro could be modified in order to offer a pet alike experience with extended capabilities, such as personal assistant, planner and conversational partner. However, critical applications involving danger, urgent medical attention or physical *HRI* should be discarded of initial real reciprocal scenarios due to service robots should accomplish a limited

unique function rather than a social interaction as is suggested by Coeckelbergh [31] and in[71] related to dangerous applications of robots.

5.4.2 Edutainment and Marketing

Long time ago Arthur C. Clarke said: Any teacher that can be replaced by a machine should be! [30]. We are still far way from this scenario; however, I think that social robots would be useful tools in the classroom. Recently several studies have been done with social robots teaching languages [77] and handwriting[69]. The handwriting example is an interesting approach using basic reciprocity in the form of feedback between the robot and the child. The child teaches the robot how to write and correct the robot telling it how to improve the traces. During the teaching process, the child learn how to do the appropriate handwriting.

I think that an approach such as the edutainment [40] using robots instead of videos or virtual characters could be more interesting if the robot were to display reciprocal interactions with the user. The robot can be a valuable support tool for the teacher if the concepts of reciprocity, gamification and edutainment are embedded in the interaction. For instance, a practice robots that requires human inputs to display rewarding educative outputs could be ideal for several educative activities.

Using the same concepts of reciprocal strategies, gamification and edutainment robots could be a powerful marketing tool. In Auckland, New Zealand it is very common interact in the street with people who ask for donations to different organisations. I think NGOs and Charities could take advantage of social robots that can perform better than humans collectors. The possibility of different attractive robot embodiments, with physical capabilities superior to the humans (able to perform attractive dances, multimedia displays and Internet access, for example) and a great ratio of cost/benefit could make possible better persuaders than the regular humans in the streets asking for donations. Some attempts like DONA[104] have been made at the moment and again, I think the use of more reciprocal robot behaviours could be implemented.

5.4.3 Training of complex social behaviours

It is likely that the most futuristic scenario in the long term is the genuine reciprocal exchange of benefits between humans and robots. Contrary to other authors like [88] I doubt that humans can be fully substituted by robots permanently. Humans require the reciprocal response of their peers and other social agents [94]. Simultaneously , they also require the unexpected behaviour that makes humans likeable, as my second and third study shown.

We are far from be able to program these "predictable" unexpected and likeable behaviours. Indeed, probably we don't need them or want them for *HRI*.

However, I think that real collaboration between robots and Artificial intelligent agents would be possible in the future. Collaboration based in reciprocity could be extremely popular once we are able to design and implement complex social behaviour in robots. Indeed robot behavioural designers could have some inspiration coming from video games, particularly strategy video games, to implement these complex social behaviours in a robot platform. The machine learning, artificial intelligence and game theory required to do that has been available since several years ago. I consider that the implementation of these complex social behaviours would require big data coming from human input. Deep learning applied in social robots would require that humans train these robots for longer periods in order to obtain reasonable realistic behaviours. I think that this machine learning could not not be done through computational simulations due to the architecture of the robots requiring simultaneous training. Movement capture, speech recognition, artificial vision, intonation of the synthetic voices, artificial inferential process, and cognitive models must be trained simultaneously in order to make them work together. What would be the benefit of the humans training these social robots?

In *HHI*, we observe that parents, teachers and peers do this training in younger humans along their lives. In the interaction parents-baby the parents are stimulated by the biological programmed emotional responses of the babies. There is very subtle reciprocal process involved in this interaction. However, in the case of *HRI*, there is not particular human interest in the training of a robot. How should the robot reciprocate towards the human trainer in order to keep him/her interested? What could be the benefit for the humans who train robots? I think that we should start to design certain mechanisms in order that humans can be reciprocated emotionally and also obtain educative and entertaining rewards when they interact with robots.

5.5 Summary

In this last chapter I drawn three main conclusions presented previously:

1. Seemingly, the Norm of Reciprocity [63] under the operationalisation of the definition of Fehr and Gaechter [53] applies in HRI as studies in Chapter 2 and Chapter 4 suggest.

2. Cooperative, recognisable , but some unexpected robot behaviour is likeable to a certain degree for the users and beneficial for the robot as Chapter 3 and Chapter 4 suggest. However, these behaviours are not the most benefical HRI strategy.

3. More studies in reciprocity are required in more complex social scenarios as the *Limitation* section of each chapter suggest. The novel effect of the robot possibly affects the results of the experiments.

These conclusions have been explained in detail based in the interpretation of my results in the previous sections. Although it is out of the scope of this thesis, in the case of the third conclusion, a set speculative applications of reciprocal scenarios in HRI have been exposed in section 5.4.

Finally I would like to express that the realisation of this thesis has been an passionate voyage. I consider that this topic is fundamental to the development of social robots independently of other factors such as degree of anthropomorphism, embodiment, and aesthetics. Besides, the development of complex reciprocal algorithms of machine learning and artificial intelligence in social robots just make sense if researchers can probe that they have a direct benefit in users and maybe in robots too.I did my best to present a consistent set of studies that could contribute to increase the corpus of knowledge in *HRI*. My big hope is that somebody join me in the future exploration of new designs of reciprocal Human-Robot Interactions.

References

[1] Abbink, K., Irlenbusch, B., and Renner, E. (2002). An experimental bribery game. *Journal of Law, Economics, & Organization*, 18(2):428–454.

[2] Amazon-Inc (2015). *Amazon Echo*. Amazon Inc.

[3] Andreoni, J. and Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, 103(418):570–585.

[4] Arrow, K. J. and Intriligator, M. D. (2006). Introduction to the series. In Kolm, S.-C. and Ythier, J. M., editors, *Foundations*, volume 1 of *Handbook of the Economics of Giving, Altruism and Reciprocity*, pages vii –. Elsevier.

[5] Axelrod, R. (1980a). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, 24(1):3–25.

[6] Axelrod, R. (1980b). More effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, 24(3):379–403.

[7] Axelrod, R. M. (1984). *The evolution of cooperation*. Basic Books, New York.

[8] Bartneck, C. (2013). Robots in the theatre and the media. In *Design & Semantics of Form & Movement (DeSForM2013)*, pages 64–70. Philips.

[9] Bartneck, C., Croft, E., and Kulic, D. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81.

[10] Bartneck, C., Kanda, T., Ishiguro, H., and Hagita, N. (2007). Is the uncanny valley an uncanny cliff? In *16th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2007*, pages 368–373, Jeju, Korea. IEEE.

[11] Berra, I. (2014). An evolutionary ockham's razor to reciprocity. *Frontiers in Psychology*, 5:1258.

[12] Birx, H. (2005). *Encyclopedia of Anthropology*. SAGE Publications, 1st edition edition.

[13] Black's Law, D. (2015). *What is Bribery?* The Black's Law Dictionary.

[14] Bluefrog-Robotics (2015). *BUDDY : Your Familys Companion Robot*. Bluefrog Robotics.

[15] Bonell, M. and Meyer, O. (2015). *The Impact of Corruption on International Commercial Contracts*. Ius Comparatum - Global Studies in Comparative Law. Springer International Publishing.

[16] Boone, C., De Brabander, B., and van Witteloostuijn, A. (1999). The impact of personality on behavior in five prisoner's dilemma games. *Journal of Economic Psychology*, 20(3):343–377.

[17] Bowlby, J. (1983). *Attachment: Attachment and Loss Volume One(Basic Books Classics)*. Basic Books, 2nd edition edition.

[18] Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., and Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 708–713. IEEE.

[19] Breazeal, C. L. (2002). *Designing sociable robots*. Intelligent robots and autonomous agents. MIT Press, Cambridge, Mass.; London.

[20] Breazeal, C. L. (2004). *Designing Sociable Robots*. MIT Press.

[21] Broadbent, E., Peri, K., Kerse, N., Jayawardena, C., Kuo, I., Datta, C., and MacDonald, B. (2014). Robots in older people's homes to improve medication adherence and quality of life: A randomised cross-over trial. In Beetz, M., Johnston, B., and Williams, M.-A., editors, *Social Robotics*, volume 8755 of *Lecture Notes in Computer Science*, pages 64–73. Springer International Publishing.

[22] Broz, F. and Lehmann, H. (2013). Do we need compassion in robots? In Weiss, A., Lorenz, T., Robins, B., Everes, V., and Vincze, M., editors, *International Conference of Social Robotics Proceedings*, Taking Care of each Other: Synchronisaton and Reciprocity for Social Companion Robots, pages 15–18. Springer International Publishing.

[23] Burnell, S. J., Evans, L., and Yao, S. (1999). The ultimatum game: Optimal strategies without fairness. *Games and economic behavior*, 26(2):221–252.

[24] Carrasco, N. and Candel, M. (2005). La justicia como reciprocidad entre individuos (epicuro) frente a la justicia como finalidad comun (aristoteles). *Convivion Revista Filosofica*, 18:3–21.

[25] Carson, A., Salt, Rachel, B. G., and Moffit, M. (2016). How is your phone changing you? *ASAPScience, Youtube*.

[26] Chaudhuri, A., Sopher, B., and Strand, P. (2002). Cooperation in social dilemmas, trust and reciprocity. *Journal of Economic Psychology*, 23(2):231–249.

[27] Cialdini, R. B. (1993). *Influence: science and practice*. HarperCollinsCollegePublishers, New York, 3rd ed edition.

[28] Cillessen, A. H. and Rose, A. J. (2005). Understanding popularity in the peer system. *Current Directions in Psychological Science*, 14(2):102–105.

[29] Clark, M. L. and Ayers, M. (1993). Friendship Expectations and Friendship Evaluations: Reciprocity and Gender Effects. *Youth & Society*, 24(3):299–313.

[30] Clarke, C. A. (1980). Sir arthur's quotations | the arthur c. clarke foundation. *Clarke Foundation*.

[31] Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12(3):235–241.

[32] Cook, R., Bird, G., Lunser, G., Huck, S., and Heyes, C. (2012). Automatic imitation in a strategic context: players of rock-paper-scissors imitate opponents' gestures. *Proceedings of the Royal Society B: Biological Sciences*, 279(1729):780–786.

[33] Cormier, D., Newman, G., Nakane, M., Young, E., J., and Durocher, S. (2013). Would you do as a robot commands? an obedience study for human-robot interaction. In *n Proceedings of the First International Conference on Human-Agent Interaction, iHAI'13.*, iHAI'13, pages I–3–1. The Japanese Society of Artificial Intelligence.

[34] Dance, G. and Jackson, T. (2014). Rock-paper-scissors: You vs. the computer. *The New York Times,* 2014.

[35] Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704.

[36] Dautenhahn, K., Woods, S., Kaouri, C., Walters, M., and Werry, I. (2005). What is a robot companion - friend, assistant or butler? *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1192–1197.

[37] Davidow, B. (2012). Exploiting the neuroscience of internet addiction. *www.stoppredatorygambling.org*.

[38] Davis, W. (2013). Strategies in iterated prisoners dilemma. *http://www.iterated-prisoners-dilemma.net/prisoners-dilemma-strategies.shtml*.

[39] Dawes, R. M. and Messick, D. M. (2000). Social dilemmas. *International journal of psychology*, 35(2):111–116.

[40] De Carolis, B. and Rossano, V. (2009). A team of presentation agents for edutainment. In *Proceedings of the 8th International Conference on Interaction Design and Children*, IDC '09, pages 150–153, New York, NY, USA. ACM.

[41] DeSteno, D. (2011). A culture of bribery? *Psychology Today*.

[42] Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 9–15, New York, NY, USA. ACM.

[43] Dictionary, M. W. (2015a). *Definition of BRIBE*. Merriam Webster Dictionary Online.

[44] Dictionary, M. W. (2015b). *Likable definition. Merriam Webster Dictionary online, 2015*. Merriam Webster Dictionary Online.

[45] Dimitri, V. d. L., Scholte, R. H., Cillessen, A. H., te Nijenhuis, J., and Segers, E. (2010). Classroom ratings of likeability and popularity are related to the big five and the general factor of personality. *Journal of Research in Personality*, 44(5):669 – 672.

[46] Draper, H., Sorell, T., Bedaf, S., Syrdal, D., Gutierrez-Ruiz, C., Duclos, A., and Amirabdollahian, F. (2014). Ethical dimensions of human-robot interactions in the care of older people: Insights from 21 focus groups convened in the uk, france and the netherlands. In Beetz, M., Johnston, B., and Williams, M.-A., editors, *Social Robotics*, volume 8755 of *Lecture Notes in Computer Science*, pages 135–145. Springer International Publishing.

[47] Dreier, T. and Spiecker, I. (2012). Legal aspects of service robotics. *Poiesis & Praxis*, 9(3-4):201–217.

[48] e.V, T. I. (2014). How corrupt is your country? *Transparency International Website*.

[49] Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, 41(1):20–26.

[50] Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293 – 315.

[51] Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

[52] Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *The journal of economic perspectives*, 14(3):159–181.

[53] Fehr, E. and Gaechter, S. (1998). Reciprocity and economics: The economic implications of homo reciprocans. *European Economic Review*, 42(3–5):845–859.

[54] Fitzpatrick, J. (2009). *Resource Theory*, pages 1371–1372. SAGE Publications, Inc., 0 edition.

[55] Fogg, B. and Nass, C. (1997). How users reciprocate to computers: An experiment that demonstrates behavior change. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '97, pages 331–332, New York, NY, USA. ACM.

[56] Fogg, B. J. (1999). Persuasive technologies. *Communications of the ACM*, 42(5):26–29.

[57] Fogg, B. J. (2002). Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):5.

[58] Fogg, B. J. (2003). *Persuasive computing: technologies designed to change attitudes and behaviors*. Morgan Kaufmann; Elsevier Science, San Francisco, Calif.; Oxford.

[59] Goldstein, N. J. (2008). *Yes!: 50 scientifically proven ways to be persuasive*. Free Press, New York, 1st. hardcover ed edition.

[60] Goodrich, M. A. and Schultz, A. C. (2007). Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275.

[61] Gosling, S. D., Rentfrow, P. J., and Swann Jr., W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

[62] Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., and Maisonnier, B. (2009). Mechatronic design of NAO humanoid. In *IEEE International Conference on Robotics and Automation, 2009. ICRA '09*, pages 769–774.

[63] Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2):161–178.

[64] Hirsh, J. B. and Peterson, J. B. (2009). Extraversion, neuroticism, and the prisoner's dilemma. *Personality and Individual Differences*, 46(2):254–256.

[65] Ho, C. and Jackson, J. W. (2001). Attitude toward asian americans: Theory and measurement. *Journal of Applied Social Psychology*, 31(8):1553–1581.

[66] Ho, C.-C. and MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, 26(6):1508 – 1518. Online Interactivity: Role of Technology in Behavior Change.

[67] Hoffman, E., McCabe, K. A., and Smith, V. L. (1998). Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology. *Economic Inquiry*, 36(3):335.

[68] Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., Hochendoner, E., and Finkenaur, J. (2015). Robot presence and human honesty: Experimental evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 181–188, New York, NY, USA. ACM.

[69] Hood, D., Lemaignan, S., and Dillenbourg, P. (2015). When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 83–90, New York, NY, USA. ACM.

[70] Jibo, I. (2015). *JIBO, The Worlds First Social Robot for the Home | Indiegogo*. Jibo Inc.

[71] Johnson, A. M. and Axinn, S. (2014). Acting vs. being moral: The limits of technological moral actors. In *Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on*, pages 1–4.

[72] Kagel, J. H. and Roth, A. E. (1995). *The handbook of experimental economics*. Princeton University Press, Princeton, N.J.

[73] Kahn, P., Ishiguro, H., Friedman, B., and Kanda, T. (2006). What is a human? - toward psychological benchmarks in the field of human-robot interaction. In *Proceedings of The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 364–371.

[74] Kahn, Jr., P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N. G., and Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, pages 33–40, New York, NY, USA. ACM.

[75] Kahn, Jr., P. H., Kanda, T., Ishiguro, H., Gill, B. T., Shen, S., Gary, H. E., and Ruckert, J. H. (2015). Will people keep the secret of a humanoid robot?: Psychological intimacy in hri. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 173–180, New York, NY, USA. ACM.

[76] Kahn Jr., P. H., Friedman, B., Perez-Granados, D. R., and Freier, N. G. (2006). Robotic pets in the lives of preschool children. *Interaction Studies*, 7(3):405–436.

[77] Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, 19(1-2):61–84.

[78] Katsuki, Y., Yamakawa, Y., and Ishikawa, M. (2015). High-speed human / robot hand interaction system. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, pages 117–118, New York, NY, USA. ACM.

[79] Kiesler, S., Sproull, L., and Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology*, 70(1):47 – 65.

[80] Kolm, S.-C. (2006a). Chapter 6 reciprocity: Its scope, rationales, and consequences. In Serge-Christophe Kolm and Jean Mercier Ythier, editor, *Handbook of the Economics of Giving, Altruism and Reciprocity*, volume Volume 1, pages 371–541. Elsevier.

[81] Kolm, S.-C. (2006b). Reciprocity: its scope, rationales, and consequences. *Handbook of the economics of giving, altruism and reciprocity*, 1:371–541.

[82] Kolm, S.-C. and Ythier, J. M. (2006). *Handbook of the economics of giving, altruism and reciprocity: Foundations*, volume 1. Elsevier.

[83] Konok, V., Gigler, D., Bereczky, B. M., and Ádám Miklósi (2016). Humans' attachment to their mobile phones and its relationship with interpersonal attachment style. *Computers in Human Behavior*, 61:537 – 547.

[84] Kreps, D. M., Milgrom, P., Roberts, J., and Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245–252.

[85] Kunz, P. R. (1969). *Romantic Love and Reciprocity*. National Council on Family Relations.

[86] Lambsdorff, J. G. and Frank, B. (2010). Bribing versus gift giving. an experiment. *Journal of Economic Psychology*, 31(3):347–357.

[87] Lammer, L., Huber, A., Weiss, A., and Vincze, M. (2014). Mutual care: How older adults react when they should help their care robot. In *AISB2014: Proceedings of the 3rd international symposium on New Frontiers in Human-Robot interaction*.

[88] Levy, D. (2009). *Love and Sex with Robots*. HarperCollins e-books.

[89] Lin, R. and Kraus, S. (2010). Can automated agents proficiently negotiate with humans? *Communications of the ACM*, 53(1):78.

[90] Litoiu, A., Ullman, D., Kim, J., and Scassellati, B. (2015a). Evidence that robots trigger a cheating detector in humans. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 165–172, New York, NY, USA. ACM.

[91] Litoiu, A., Ullman, D., Kim, J., and Scassellati, B. (2015b). Evidence that robots trigger a cheating detector in humans. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 165–172, New York, NY, USA. ACM.

[92] Lorenz, T. (2013). Synchrony and reciprocity for social companion robots: benefits and challenges. In Weiss, A., Lorenz, T., Robins, B., Everes, V., and Vincze, M., editors, *International Conference of Social Robotics Proceedings*, Taking Care of each Other: Synchronisaton and Reciprocity for Social Companion Robots, pages 10–14. Springer International Publishing.

[93] Luke (2016). Passage Luke, 6:35. *Bible Gateway passage Online*.

[94] Mail, B. C. F. S. C. F. T. D. (2016). How robots will soon take teens' virginity. *DailyMail*.

[95] Malinowski, B. (1961). *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. E. P. Dutton & Co.

[96] Malle, B., Scheutz, M., and Austerweil, J. (In press). *A world with robots Chapter: Networks of Social and Moral Norms in Human and Robot Agents*. Springer Verlag.

[97] Malle, B. F. (2014). Moral competence in robots? *Sociable robots and the future of social relations: Proceedings of Robo-Philosophy 2014*, pages 189–198.

[98] Mann, J. A., MacDonald, B. A., Kuo, I.-H., Li, X., and Broadbent, E. (2015). People respond better to robots than computer tablets delivering healthcare instructions. *Computers in Human Behavior*, 43:112 – 117.

[99] Marino, F. (1999-2009). Murder on the planet express, futurama.

[100] Matthew (2016). Passage Matthew, 7:12. *Bible Gateway passage Online*.

[101] McCall, C., Maynes, B., Zou, C. C., and Zhang, N. J. (2013). An automatic medication self-management and monitoring system for independently living patients. *Medical engineering & physics*, 35(4):505–514.

[102] Melo, C. M. D., Zheng, L., and Gratch, J. (2009). Expression of Moral Emotions in Cooperating Agents *. *Intelligent Virtual Agents*.

[103] Mervielde, I. and De Fruyt, F. (2000). The big five personality factors as a model for the structure of children's peer nominations. *European Journal of Personality*, 14(2):91–106.

[104] Min Su, K., Dong Min, P., Byung Keun, C., Sae Mee, L., Sonya, K., and Min Kyung, L. (2011). DONA - plastic pals. *plasticpals.com*.

[105] Molm, L. D. (2010). The structure of reciprocity. *Social Psychology Quarterly*, 73(2):119–131.

[106] Mori, M., MacDorman, K., and Kageki, N. (2012). The uncanny valley [from the field]. *Robotics Automation Magazine, IEEE*, 19(2):98–100.

[107] Mumm, J. and Mutlu, B. (2011). Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 331–338. ACM.

[108] Muscolo, G. G., Recchiuto, C. T., Campatelli, G., and Molfino, R. (2013). A robotic social reciprocity in children with autism spectrum disorder. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8239 LNAI, pages 574–575.

[109] Myers, D. G. (2009). *Social Psychology, 10th Edition*. McGraw-Hill, 10th edition.

[110] Nass, C. and Moon, Y. (2000a). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1):81–103.

[111] Nass, C. and Moon, Y. (2000b). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1):81–103.

[112] Nass, C. and Reeves, B. (1996). *The media equation: how people treat computers, televisions, and new media like real people and places*. CSLI Publications; Cambridge University Press, Stanford, Calif.: New York; Cambridge.

[113] Nishio, S., Ogawa, K., Kanakogi, Y., Itakura, S., and Ishiguro, H. (2012a). Do robot appearance and speech affect people's attitude? evaluation through the ultimatum game. In *2012 IEEE RO-MAN*, pages 809–814.

[114] Nishio, S., Ogawa, K., Kanakogi, Y., Itakura, S., and Ishiguro, H. (2012b). Do robot appearance and speech affect people's attitude? evaluation through the ultimatum game. In *RO-MAN, 2012 IEEE*, pages 809–814. IEEE.

[115] Norman, D. (2009). *The Design of Future Things*. Basic Books, first trade paper edition edition.

[116] Norman, D. A. (2002). *The Design of Everyday Things*. Basic Books, first edit edition.

[117] Norman, D. A. (2005). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, 1 edition.

[118] Oda, R. (1997). Biased face recognition in the prisoner's dilemma game. *Evolution and Human Behavior*, 18(5):309–315.

[119] Oosterbeek, H., Sloof, R., and Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188.

[120] Packard, V. and Miller, M. C. (2007). *The Hidden Persuaders*, chapter Introduction. Ig Publishing, reissue edition edition.

[121] Parisi, F. (2003). Reciprocity. *The Encyclopedia of Public Choice*, pages 797–802.

[122] Park, H. and Antonioni, D. (2007). Personality, reciprocity, and strength of conflict resolution strategy. *Journal of Research in Personality*, 41(1):110–125.

[123] Parliament of the World's Religions (2016). Declaration toward a global ethic. *Parliament of the World's Religions Website*.

[124] Perugini, M., Gallucci, M., Presaghi, F., and Ercolani, A. P. (2003). The personal norm of reciprocity. *European Journal of Personality*, 17(4):251–283.

[125] Pinker, S., Nowak, M. A., and Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3):833–838.

[126] Polcino, M. and Anderson, M. B. (2012). Them, robot. the simpsons.

[127] Poundstone, W. (2011). *Prisoner's Dilemma*. Anchor.

[128] R, K. (2013). Game Theory: Assumptions, Application and Limitations. *Business Management Ideas*.

[129] Rapoport, A. (1965). *Prisoner's dilemma: a study in conflict and cooperation*. University of Michigan press, Ann Arbor, Mich.

[130] Robinson, H., MacDonald, B., and Broadbent, E. (2015). Physiological effects of a companion robot on blood pressure of older people in residential care facility: A pilot study. *Australasian Journal on Ageing*, 34(1):27–32.

[131] Robinson, H., MacDonald, B., Kerse, N., and Broadbent, E. (2013). The psychosocial effects of a companion robot: A randomized controlled trial. *Journal of the American Medical Directors Association*, 14(9):661–667.

[132] Roizman, M., Hoffman, G., Ayal, S., Hochman, G., Tagar, M. R., and Maaravi, Y. (2016). Studying the opposing effects of robot presence on human corruption. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 501–502.

[133] Ross, S. M. (2003). Peirce's criterion for the elimination of suspect experimental data. *Journal of Engineering Technology*.

[134] Sahlins, M. (2003). *Stone Age Economics*. Routledge, 2 edition.

[135] Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 141–148, New York, NY, USA. ACM.

[136] Sandoval, E. B., Brandstetter, J., and Bartneck, C. (2016a). Can a robot bribe a human?: The measurement of the negative side of reciprocity in human robot interaction. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pages 117–124, Piscataway, NJ, USA. IEEE Press.

[137] Sandoval, E. B., Brandstetter, J., Obaid, M., and Bartneck, C. (2016b). Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics*, 8(2):303–317.

[138] Sandoval, E. B., Brandstetter, J., and Utku Bartneck, C. (In submission). Measurement of reciprocal strategies in human robot interaction and their likeability using the alternated repeated ultimatum game. In *International Journal of Social Robotics*, IJSR.

[139] Scheutz, M. and Arnold, T. (2016). Are we ready for sex robots? In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pages 351–358, Piscataway, NJ, USA. IEEE Press.

[140] Schreier, J. (2012). Robot & frank movie.

[141] Selten, R. and Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames a learning theory approach. *Journal of Economic Behavior & Organization*, 7(1):47 – 70.

[142] Short, E., Hart, J., Vu, M., and Scassellati, B. (2010). No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 219–226.

[143] Slembeck, T. (1999). Reputations and fairness in bargaining-experimental evidence from a repeated ultimatum game with fixed opponents. Technical report, EconWPA.

[144] Softbank-Aldebaran (2015). *Who is Pepper?* Softbank Robotics.

[145] Solnick, S. J. and Schweitzer, M. E. (1999). The influence of physical attractiveness and gender on ultimatum game decisions. *Organizational behavior and human decision processes*, 79(3):199–215.

[146] Sophister, L. D.-J. (2000). Public goods and the prisoner's dilemma: Experimental evidence. *Student Economic Review*.

[147] Spaniel, W. (2011). *Game Theory 101: The Complete Textbook*. Amazon.

[148] Sullivan, L. E. (2009). Lex talionis. *In The SAGE glossary of the social and behavioral sciences*, 3:290–290.

[149] Torta, E., van Dijk, E., Ruijten, P. A. M., and Cuijpers, R. H. (2013). *Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings*, chapter The Ultimatum Game as Measurement Tool for Anthropomorphism in Human–Robot Interaction, pages 209–217. Springer International Publishing, Cham.

[150] Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly review of biology*, pages 35–57.

[151] Turkle, S. (2011). *Alone together: why we expect more from technology and less from each other*. Basic Books, New York.

[152] Ullman, D. and Malle, B. (2016). The effect of perceived involvement on trust in human-robot interaction. In *The Eleventh ACM/IEEE International Conference on Human Robot Interation*, pages 641–642. IEEE Press.

[153] Voiklis, J., K. B. C. C. and Malle, B. ((2016, August)). Moral judgments of human vs. robot agents. In *In Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016).*

[154] Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology*, 12(3):243–250.

[155] Weiss, A. (2013). Grounding in human-robot interaction: Can it be achieved with the help of the user? In Weiss, A., Lorenz, T., Robins, B., Everes, V., and Vincze, M., editors, *International Conference of Social Robotics Proceedings*, Taking Care of each Other: Synchronisaton and Reciprocity for Social Companion Robots, pages 7–10. Springer International Publishing.

[156] Weiss, A. and Lorenz, T. (2013). Icsr 2013 workshop 3: Final report and results. taking care of each other:sincronization and reciprocity for social companion robots. In Weiss, A., Lorenz, T., Robins, B., Everes, V., and Vincze, M., editors, *International Conference of Social Robotics 2013 Proceedings*, Taking Care of each Other: Synchronisaton and Reciprocity for Social Companion Robots, pages 1–7. Springer International Publishing.

[157] Weiss, A. and Tscheligi, M. (2010). Special issue on robots for future societies: evaluating social acceptance and societal impact of robots. *International Journal of Social Robotics*, 2(4):345–346.

[158] Whatley, M. A., Webster, J. M., Smith, R. H., and Rhodes, A. (1999). The effect of a favor on public and private compliance: How internalized is the norm of reciprocity? *Basic and Applied Social Psychology*, 21(3):251–259.

[159] Williams, K. C. (2012). *Introduction to Game Theory: A Behavioral Approach*. Oxford University Press, USA.

[160] Wilson, R. H. and Stoy, M. A. (2001). Automated medication-dispensing cart. US Patent 6,170,929.

[161] Yildirim, C. and Correia, A.-P. (2015). Exploring the dimensions of nomophobia: Development and validation of a self-reported questionnaire. *Computers in Human Behavior*, 49:130 – 137.

[162] Zefferman, M. R. (2014). Direct reciprocity under uncertainty does not explain one-shot cooperation, but demonstrates the benefits of a norm psychology. *Evolution and Human Behavior*, 35(5):358 – 367.

Appendix A

Outcomes in the PhD

Further information about the outcomes of my PhD are available in www.sandoval.nz

A.1 Full papers

- E. B. Sandoval, J. Brandstetter, U, Yalcin, C. Bartneck "Measurement of Reciprocal Strategies in Human Robot Interaction and their likeability using the Alternated Repeated Ultimatum Game," Human-Agent Interaction (HAI), 2016 4th ACM/IEEE International Conference on, 2016, pp XX-XX.
- E. B. Sandoval, J. Brandstetter, C. Bartneck "Can a Robot Bribe a Human? The Measurement of the Negative Side of Reciprocity in Human Robot Interaction," Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, 2016, pp 117-124.
- E. B. Sandoval, J. Brandstetter, M. Obaid, and C. Bartneck, "Reciprocity in Human-Robot Interaction: A Quantitative Approach through the Prisoner's Dilemma and the Ultimatum Game," International Journal of Social Robotics, 2015, pp. 1-15.
- E. B. Sandoval and O. Mubin, "Making HRI accessible to everyone through online videos: A proposal for a microMOOC in human robot interaction," Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on, Kobe, 2015, pp60-64.
- E.B. Sandoval, O. Mubin, and M. Obaid. "Human Robot Interaction and Fiction: A Contradiction," Social Robotics, 6th International Conference, ICSR 2014, Sydney, NSW, Australia, October 27-29, 2014. Proceedings, M. Beetz, B. Johnston, and M.-A. Williams, Eds. Cham: Springer International Publishing, 2014. 54-63.

- M. Obaid, E.B. Sandoval, J. Złotowski, E. Moltchanova, C.A. Basedow and C. Bartneck, "Stop! That is Close Enough. How Body Postures Influence Human-Robot Proximity," Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on, New York, 2016, pp XXX-XXX.

- O. Mubin, M. Obaid, E. Sandoval, and M. Fjeld, "Using Video Preferences to Understand the Human Perception of Real and Fictional Robots," in Proceedings of the 3rd International Conference on Human-Agent Interaction, New York, NY, USA, 2015, pp. 33–39.

- C. Bartneck, M. Soucy, K. Fleuret and E. B. Sandoval, "The robot engine — Making the unity 3D game engine work for HRI," Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on, Kobe, 2015, pp. 431-437.

- J. Brandstetter, P. Rácz, C. Beckner, E. B. Sandoval, J. Hay and C. Bartneck, "A peer pressure experiment: Recreation of the Asch conformity experiment with robots," Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on, Chicago, IL, 2014, pp. 1335-1340.

A.2 Short papers

- E.B. Sandoval, O.Rudhru, Q. Min Ser The Birth of a New Discipline: Robotology. A First Robotologist Study over a Robot Maori Haka. Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on,Christchurch, New Zealand, 2016, pp 511-512.

- Omprakash Rudhru, Qi Min Ser, and Eduardo Benitez Sandoval. 2016. Robot Maori Haka: Robots as cultural preservationists. In The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16). IEEE Press, Piscataway, NJ, USA, 569-569. VIDEO

- Qi Min Ser, Omprakash Rudhru, and Eduardo Benitez Sandoval. 2016. Robot Maori Haka. In The Eleventh ACM/IEEE International Conference on Human Robot Interation (HRI '16). IEEE Press, Piscataway, NJ, USA, 549-549. DEMO

- E.B. Sandoval, J. Brandstetter. 2015, MicroMOOCs for HRI teaching. Workshop of education in HRI. Human-Robot Interaction (HRI), 2015 10th ACM/IEEE International Conference on Portland, USA, 2016. Not published in the Proceedings

A.3 Outreach articles

- O. Mubin and E. B. Sandoval, "De la science-fiction aux nouvelles technologies, et vice et versa," The Conversation, 06-Jan-2016.
- O. Mubin and E. B. Sandoval, "The technology in science fiction is not always what we want in the real world," The Conversation, 04-Dec-2015.

A.4 Videos and Demos

- E. B. Sandoval, O. Rudhru, and Q. Min Ser, Robot Maori Haka: Robots as cultural preservationists, Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, 2016. We won the ACM/IEEE **1st place HRI video Award, HRI Conference 2016**
- O. Rudhru, and Q. Min Ser,E. B. Sandoval, Nao Robot Haka 2016 Demo, Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, 2016, Youtube
- Eduardo B. Sandoval, Robot bribes people, Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, 2016,
- Eduardo B. Sandoval, Making Human Robot Interaction Accessible To Everyone Through Online Videos, Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on, Kobe, 2015,
- Eduardo B. Sandoval, Robot Oracle predicts the Winner of Cricket Worldcup 2015, HITLabNZ, 2015.
- Eduardo B. Sandoval and O. Mubin, What is Human Robot Interaction? HITLabNZ, 2015,
- Eduardo B. Sandoval, M. Obaid, and O. Mubin, Fiction robots vs real robots, HIT-LabNZ 2014.
- E. B. Sandoval and M. Adair, Eugene in Aotearoa. Robots and Art – Misbehaving Machines workshop, Social Robotics: 6th International Conference, ICSR 2014, Sydney, NSW, Australia, October 27-29, 2014.
- Eduardo B. Sandoval and S. Sohnchen, Design Thinking University of Canterbury. Design Thinking Workshop HITLabNZ, 2013.

- Bell, Michael and E. B. Sandoval, Computer Science Field Guide: Artificial Intelligence, Youtube csunplugged/ Orange Studio, 2013.
- E. B. Sandoval, Nao Robot Dances Gangnam Style. HITLabNZ, 2012. Almost 1,000,000 million views!

A.5 Awards

- 1st Place Video HRI Award, Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, 2016, Christchurch, New Zealand, 2016.
- 30 promising young mexicans, CNN Expansion Magazine, Mexico, 2015.
- Medal Alfonso Caso 2014, National Autonomous University of Mexico, best student graduated in 2012 of the Master of Industrial Design program, 128 awards for a total population of 28,000 graduate students, Mexico, September, 2014.
- Honorific Mention, Startup Weekend Christchurch, project FoodFlow, March 2014.
- Audience Award for the Model of Reciprocity in HRI, Showcase 2012, University of Canterbury Postgraduate Student Association, 2012.
- NEC NZ Ltd Scholarship, 2012-2015.
- UC Doctoral Scholarship, 2013-2016.
- Conacyt International PhD Scholarship, 2013-2016.

A.6 Presentations

-
- Selwyn Retirement house, Reciprocity in Human-Robot Interaction, 30th March 2016
- Futureintech Program, Liston College, Can a robot bribe a human? 24th January 2016.
- Futureintech Program, Christchurch Cathedral Grammar. Life when you are twice as old, 18th Nov 2015
- Kshitij 2014 Festival, Indian Institute of Technology, Social Robots and their impact, **Kharagpur, India, 2014.**

- IPENZ, Futureintech is an IPENZ initiative, Christchurch, New Zealand 2014.
- Ministry of Inspiration, Nelson New Zealand 2013.
- Showcase University of Canterbury 2012, 2014, 2015
- Several other small presentations for schools in Christchurch and Auckland.

A.7 Volunteering

- I am mentor of TechInFuture of IPENZ for Epsom Grammar Girls School, Auckland.
- I have been reviewer in conferences like: HRI'16 , ICSR'15 , HAI-14, HAI-16, DIS'16 (Designing for interactive systems), IDC'16 (Interaction design for children), NordiCHI 2016, and Journal of Sensors and Actuators A. Physical.
- IPENZ, FutureInTech ambassador, 2014-2016.
- Student's representative in the HITLab NZ, 2013-2014.
- Red Global de Mexicanos, Commite member, 2013/2015.
- Immigration Centre Christchurch, I have given talks about Digital skills for business, 2014/2015.
- TEDx Christchurch, Volunteer, 2013-2015.
- Funder of the UC Spanish club.
- Canterbury-Westland Schools, Science and Technology Fair, 2012, NZ.
- About 40 demos and presentations of Human-Robot Interaction since 2012 in NZ.

A.8 Teaching

- Auckland University of Technology, Colab, Social Robotics Methods, 21th Sept 2016.
- Auckland University of Technology, Colab, How to Program the NAO robots, April 2016.
- University of Canterbury, Human Interface Technology Lab NZ, I have teach:How to Program the NAO robots (2013, 2014, 2015). Tutor in the Design Thinking lecture (2013) and Robot Puppetry (2013).